## MITOCW | watch?v=phbw9r1iUDI

So welcome back. So we are now moving to a new chapter, which is going to have a little more of a statistical flavor when it comes to designing methods, all right? Because if you think about it, OK-- some of you have probably attempted problem number two in the problem set. And you realize that maximum likelihood estimators does not give you super trivial estimators, right? I mean, when you have an n theta theta, then the thing you get is not something you could have guessed before you actually attempted to solve that problem. And so, in a way, we've seen already sophisticated methods.

However, in many instances, the maximum likelihood estimator was just an average. And in a way, even if we had this confirmation for maximum likelihood that indeed that was the estimator that maximum likelihood would spit out, and that our intuition was therefore pretty good, most of the statistical analysis or use of the central limit theorems, all these things actually did not come in the building of estimator, in the design of the estimator, but really in the analysis of the estimator. And you could say, well, if I know already that the best estimator is the average, I'm just going to use the average. I don't have to, basically, quantify how good it is. I just know it's the best I can do.

We're going to talk about tests. And we're going to talk about parametric hypothesis testing. So you should view this as-- parametric means, well, it's about a parameter, like we did before. And hypothesis testing is on the same level as estimation. And on the same level as estimator will be the word "test," OK? And when we're going to devise a test, we're going to actually need to understand random fluctuations that arise from the central limit theorem better, OK? It's not just going to be in the analysis. It's also going to be in the design. And everything we've been doing before in understanding the behavior of an estimator is actually going to come in and be extremely useful in the actual design of tests, OK?

So as an example, I want to talk to you about some real data. I will not study this data. But this data actually exist. You can find it on R. And so, it's the data from the so-called credit union Cherry Blossom Run, which is a 10 mile race. It takes place every year in D.C. It seems that some of the years are pretty nice.

In 2009, there were about 15,000 participants. Pretty big race. And the average running time was 103.5 minutes, all right? So about an hour and a half or a little bit more.

And so, you can ask the following question, right? This is actual data, right? 103.5 actually averaged the running time for all of 15,000. Now, this in practice, may not be something very suitable. And you might want to just sample a few runners and try to understand how they're behaving every year without having to collect the entire data set.

And so, you could ask the question, well, let's say my budget is to ask for maybe 10 runners what their running time was. I still want to be able to determine whether they were running faster in 2012 than in 2009. Why do I put 2012, and not 2016? Well, because the data set for 2012 is also available. So if you are interested and you know how to use R, just go and have fun with it.

So to answer this question, what we do is we select n runners, right? So n is a moderate number that's more manageable than 15,000. From the 2012 race at random. That's where the random variable is going to come from, right? That's where we actually inject randomness in our problem.

So remember this is an experience. So really in a way, the runners are the omegas. And I'm interested in measurements on those guys. So this is how I have a random variable. And this random verbal here is measuring their running time. OK. If you look at the data set, you have all sorts of random variables you could measure about those random runners. Country of origin. I don't know, height, age, a bunch of things. OK. Here, the random variable of interest being the running time. OK. Everybody understand what the process is?

OK. So now I'm going to have to make some modeling assumptions. And here, I'm actually pretty lucky. I actually have all the data from a past year. I mean, this is not the data from 2012, which I also have, but I don't use. But I can actually use past data to try to understand what distribution do I have, right? I mean, after all, running time is going to be rounded to something. Maybe I can think of it as a discrete random variable. Maybe I can think of it as the exponential random variable. Those are positive numbers. I mean, there's many kind of running times that could come up to mind. Many kind of distributions I could think of for this modeling part.

But it turns out that if you actually plug the histogram of those running times for all 15,000 runners in 2009, you actually are pretty happy to see that it really looks like a bell-shaped curve, which suggest that this should be a Gaussian. So what you go on to do is you estimate the mean from past observations, which was actually 103.5, as we said. You submit the variance, which was 373. And you just try to superimpose the curve with this one, which is a Gaussian PDF with mean 103.5 and variants 373. And you see that they actually look very much alike. And so here, you're pretty comfortable to say that the running time actually is Gaussian distribution. All right?

So now I know that the x1 to xn, I'm going to say they're Gaussian, OK? I still need to specify two parameters. So what I want to know is, is the distribution the same from past years, right? So I want to know if the random variable that I'm looking for-- if I, say, pick one. Say, x1. Does it have the same distribution in 2012 that it did in 2009? OK.

And so, the question is, is x1 has a Gaussian distribution with mean 103.5 and variance 373? Is that clear? OK.

So this question that calls for a yes or no answer is a hypothesis testing problem. I am testing a hypothesis. And this is the basis of basically all of data-driven scientific inquiry. You just ask questions. You formulate a scientific hypothesis.

Knocking down this gene is going to cure melanoma, is this true? I'm going to collect. I'm going to try. I'm to observe some patients on which I knock down this gene. I'm going to collect some measurements. And I'm going to try to answer this yes/no question, OK? It's different from the question, what is the mean running time for this year?

OK. So this hypothesis testing is testing if this hypothesis is true. The hypothesis in common English we just said, were runners running faster? All right? Anybody could formulate this hypothesis.

Now, you go to a statistician. And he's like, oh, what you're really asking me is x1 has a Gaussian distribution with mean less than 103.5 and variance 373, right? That's really the question that you ask in statistical terms. And so, if you're asking if this was the same as before, there's many ways it could not be the same as before. There's basically three ways it could not be the same as before.

It could be the case that x1 is in expectation to 103.5 So the expectation has changed. Or the variance has changed. Or the distribution has changed. I mean, who knows? Maybe runners are now just all running holding their hands. And it's like now a point mass at 1 given point. OK. So you never know what could [INAUDIBLE]. Now of course, if you allow for any change, you will find change. And so what you have to do is to factor in as much knowledge as you can. Make as many modeling assumptions, so that you can let the data speak about your particular question.

Here, your particular question is, are they running faster? So you're only really asking a question about the expectation. You really want to know if the expectation has changed. So as far as you're concerned, you're happy to make the assumption that the rest has been unchanged. OK. And so, this is the question we're asking. Is the expectation now less than 103.5? Because you specifically asked whether runners were going faster this year, right? They tend to go faster rather than slower, all right? OK. So this is the question we're asking in mathematical terms.

So first, when I did that, I need to basically fix the rest. And fixing the rest is actually part of the modeling assumptions. So I fixed my variance to be 373. OK? I assume that the variance has not changed between 2009 and 2012. Now, this is an assumption. It turns out it's wrong. So if you look at the data from 2012, this is not the correct assumption. But I'm just going to make it right now for the sake of argument, OK? And also the fact that it's Gaussian.

Now, this is going to be hard to violate, right? I mean, where did this bell-shaped curve come from? Well, it's just natural when you just measure a bunch of things. The central limit theorem appears in the small things of nature. I mean, that's the bedtime story you get about the central limit theorem. And that's why the bell-shaped curve is everywhere in nature. It's the sum of little independent things that are going on. And this Gaussian assumption, even if I wanted to relax it, there's not much else I can do. It is pretty robust across the years.

All right. So the only thing that we did not fix is the expectation of x1, which now I want to know what it is. And since I don't know what it is, I'm going to call it mu. And it's going to be a variable of interest, all right? So it's just a number mu. Whatever this is I can try to estimate it, maybe using maximum likelihood estimation. Probably using the average, because this is Gaussian. And we know that the maximum likelihood estimator for a Gaussian is just the average. And now we only want to test if mu is equal to 103.5, like it was in 2009. Or on the contrary, if mu is not equal to 103.5. And more specifically, if mu is actually strictly less than 103.5. That's the question you ask.

Now, why am I in writing mu equal to 103.5 or is less than 103.5 and equal to 103.5 versus not equal to 103.5? It's because since you asked me a more precise question, I'm going to be able to give you a more precise answer. And so, if your question is very specific-- are they running faster? I'm going to factor that in what I write. If you just ask me, is it the same? I'm going to have to write, or is it different than 103.5? And that's less information about what you're looking for, OK?

So by making all these modeling assumptions-- the fact that the variance doesn't change, the fact that it's still Gaussian-- I've actually reduced the number of. And I put numbers in quotes, because this is still an infinite of them. But I'm limiting the number of ways the hypothesis can be violated. The number of possible alternative realities for this hypothesis, all right?

For example, I'm saying there's no way mu can be larger than 103.5. I've already factored that in, OK? It could be. But I'm actually just going to say that if it's larger, all I'm going to be able to tell you is that it's not smaller. I'm not going to be able to tell you that it's actually larger, OK?

And the only way it can be rejected now. The only way I can reject my hypothesis is if x belongs to very specific family of distributions. If it has a distribution which is Gaussian with mean mu and variance of 373 for mu, which is less 103.5. All right?

So we started with basically was x1-- so that's the reality. x1 follows n 103.5 373, OK? And this is everything else, right? So for example, here is x follows some exponential, 0.1, OK? This is just another distribution here. Those are all the possible distributions.

What we said is we said, OK, first of all, let's just keep only those Gaussian distributions, right? And second, we said, well, among those Gaussian distributions, let's only look at those that have-- well, maybe this one should be at the boundary-- let's only look at the Gaussians here. So this guy here are all the Gaussians with mean mu and variance 373 for mu less than 103.5, OK?

So when you're going to give me data, I'm going to be able to say, well, am I this guy? Or am I one of those guys? Rather than searching through everything. And the more you search the easier for you to find something that fits better the data, right? And so, if I allow everything possible, then there's going to be something that just by pure randomness is actually going to look better for the data, OK?

So for example, if I draw 10 random variables, right? If n is equal to 10. And let's say they take 10 different values. Then it's actually more likely that those guys come from a discrete distribution that takes each of these values with probability 1 over 10, than actually some Gaussian random variable, right? That would be perfect. I can actually explain it.

If the 10 numbers I got were say-- let's say I collect 3, 90, 95, and 102. Then the most likely distribution for those guys is the discrete distribution that takes three values, 91 with probability 1/3, 95 with probability 1/3, and 102 with probably 1/3, right? That's definitely the most likely distribution for this. So if I allowed this, I would say, oh no. This is not distributed according to that. It's distributed according to this very specific distribution, which is somewhere in the realm of all possible distributions, OK?

So now we're just going to try to carve out all this stuff by making our assumptions. OK. So here in this particular example, just make a mental note that what we're doing is that I actually-- a little birdie told me that the reference number is 103.5, OK? That was the thing I'm actually looking for. In practice, it's actually seldom the case that you have this reference for yourself to think of, right? Maybe here, I just happen to have a full data set of all the runners of 2009. But if I really just asked you, I said, were runners faster in 2012 than in 2009? Here's $10 to perform your statistical analysis. What you're probably going to do is called maybe 10 runners from 2012, maybe 15 runners from 2009, ask them and try to compare their mean. There's no standard reference. You would not be able to come up with this 103.5, because these data maybe is expensive to get or something.

OK. So this is really more of the standard case, all right? Where you really compare two things with each other, but there's no actual ground truth number that you're comparing it to. OK. So we'll come back to that in a second. I'll tell you what the other example looks like.

So let's just stick to this example. I tell you it's 103.5, OK? Let's try to have our intuition work the same way. We said, well, averages worked well. The average, tell me, of over these 10 guys should tell me what the mean should be. So I can just say, well x bar is going to be close to the true mean by the law of large number. So I'm going to decide whether x bar is less than 103.5. And conclude that in this case, indeed mu is less than 103.5, because those two quantities are close, right? I could do that.

The problem is that this could go pretty wrong. Because if n is small, then I know that xn bar is not equal to mu. I know that xn bar is close to mu. But I also know that there's pretty high chance that it's not equal to mu. In particular, I know it's going to be somewhere at 1 over root n away from mu, right? 1 over root n being the root coming from what? CLT, right? That's the root n that comes from CLT. In blunt words, CLT tells me the mean is at distance 1 over root n from the expectation, pretty much. That's what it's telling.

So 1 over root n. If I have 10 people in there, 1 over root 10 is not a huge number, right? It's like 1/3 pretty much. So 1/3 103.5. If the true mean was actually 103.4, but my average was telling me it's 103.4 plus 1/3, I would actually come to two different conclusions, right?

So let's say that mu is equal to 103.4, OK? So you're not supposed to know this, right? That's the hidden truth. OK.

Now I have n is equal to 10. So I know that x bar n minus 103.4 is something of the order of 1 over the square root of 10, which is of the order of, say, 0.3. OK. So here, this is all hand wavy, OK? But that's what the central limit theorem tells me.

What it means is that it is possible that x bar n is actually equal to is actually equal to 103.4 plus 0.3, which is equal to 103.7. Which means that while the truth is that mu is less than 103.5, then I would conclude that mu is larger than 103.5, OK? And that's because I have not been very cautious, OK?

So what we want to do is to have a little buffer to account for the fact that xn bar is not a precise value for the true mu. It's something that's 1 over root n away from you. And so, what we want is the better heuristic that says, well, if I want to conclude that I'm less than 103.5, maybe I need to be less than 103.5 minus a little buffer that goes to 0 as my sample size goes to infinity. And actually, that's what the law of large number tells me. The central limit theorem actually tells me that this should be true, something that goes to 0 as n goes to infinity and the rate 1 over root n, right? That's basically what the central limit theorem tells me.

So to make this intuition more precise, we need to understand those fluctuations. We need to actually put in something that's more precise than these little wiggles here, OK? We need to actually have the central limit theorem come in.

So here is the example of comparing two groups. So pharmaceutical companies use hypothesis testing to test if a drug is efficient, right? That's what they do. They want to know, does my new drug work? And that's what the Federal Drug Administration office is doing on a daily basis. They ask for extremely well regulated clinical trials on a thousand people, and check, does this drug make a difference? Did everybody die? Does it make no difference? Should people pay $200 for a pill of sugar, right? So that's what people are actually asking.

So to do so, of course, there is no ground truth about-- so there's actually a placebo effect. So it's not like actually giving a drug that does not work is going to have no effect on patients. It will have a small effect, but it's very hard to quantify. We know that it's there, but we don't know what it is. And so rather than saying, oh the ground truth is no improvement, the ground truth is the placebo effect. And we need to measure what the placebo effect is.

So what we're going to do is we're going to split our patients into two groups. And there's going to be what's called a test group and a control group. So the word test here is used in a different way than hypothesis testing. So we'll just call it typically the drug group. And so, I will refer to mu drug for this guy, OK?

Now, this let's say this is a cough syrup, OK? And when you have a cough syrup, the way you measure the efficacy of a cough syrup is to measure how many times you cough per minute, OK? And so, if I define mu control the number of expectoration per hour. So just the expected number, right? This is the number I don't know, because I don't have access to the entire population of people that will ever take this cough syrup.

And so, I will call it mu control for the control group. So those are the people who have been actually given just like sugar, like maple syrup. And mu drug are those people who are given the actual syrup, OK? And you can imagine that maybe maple syrup will have an effect on expectorations per hour just because, well, it's just sweet and it helps, OK? And so, we don't know what this effect is going to be. We just want to measure if the drug is actually having just a better impact on expectorations per hour than the just pure maple syrup, OK?

So what we want to know is if mu drug is less than mu control. That would be enough. If we had access to all the populations that will ever take the syrup for all ages, then we would just measure, did this have an impact? And even if it's a slightly ever so small impact, then it's good to release this cough syrup, assuming that it has no side effects or anything like this, because it's just better than maple syrup, OK? The problem is that we don't have access to this. And we're going to have to make this decision based on samples that give me imprecise knowledge about mu drug and mu control.

So in this case, unlike the first case where we compared an unknown expected value to have a fixed number, which was one of the 103.5, here, we're just comparing two unknown numbers with each other, OK? So there's two sources of randomness. Trying to estimate the first one. And trying to estimate the second one.

Before I move on, I just wanted to tell you I apologize. One of the graders was not able to finish grading his problem sets for today. So for those of you who are here just to pick up their homework, feel free to leave now. Even if you have a name tag, I will pretend I did not read it. OK. So I'm sorry. You'll get it on Tuesday. And this will not happen again. OK.

So for the clinical trials, now I'm going to collect information. I'm going to collect the data from the control group. And I'm going to collect data from the test group, all right?

So my control group here. I don't have to collect the same number of people in the control group than in the drug group. Actually, for cough syrup, maybe it's not that important. But you can imagine that if you think you have the cure to a really annoying disease, it's actually hard to tell half of the people you will get a pill of nothing, OK? People tend to want to try the drug. They're desperate. And so, you have to have this sort of imbalance between who is getting the drug and who's not getting the drug.

And people have to qualify for the clinical trials. There's lots of fluctuations that affect what the final numbers of people who are actually going to get the drug and are going to get the control is going to be. And so, it's not easy for you to make those two numbers equal. You'd like to have those numbers equal if you can, but not necessarily. And by the way, this is all part of some mystical science called "design of experiments." And in particular, you can imagine that if one of the series had higher variants, you would want to like more people in this group than the other group. Yeah?

**STUDENT:** So when we're subtracting [INAUDIBLE] something that [INAUDIBLE] 0 [INAUDIBLE] to be satisfied. So that's on purpose [INAUDIBLE].

**PROFESSOR:** Yeah, that's on purpose. And I'll come to that in a second, all right? So basically, we're going to make it if your answer is, is this true? We're going to make it as hard as possible, but no harder for you to say yes to this answer. Because, well, we'll see why.

OK, so now we have two set of data, the x's and the y's. The x's are the ones for the drug. And the y's are the data that I collected from the people, who were just given a placebo, OK? And so, they're all IID random variables. And here, since it's the number of expectorations, I'm making a blunt modeling assumption. I'm just going to say it's Poisson. And it's characterized only by the mean mu drug or the mean mu control, OK? I've just made an assumption here. It could be something different. But let's say it's a Poisson distribution.

So now what I want to know is to test whether mu drug is less than mu control. We said that already. But the way we said it before was not as mathematical as it is now. Now we're actually making a test on the parameters of Poisson distribution. Whereas before, we were just making test on expected numbers, OK?

So the heuristic-- again, if we try to apply the heuristic now. Rather than comparing mu x bar drug to some fixed number, I'm actually comparing x bar drug to some control. But now here, I need to have something that accounts for, not only the fluctuations of x bar drug, but also for the fluctuations of x bar control, OK? And so, now I need something that goes to 0 when all those two things go to infinity. And typically, it should go to zero with 1 over root of n drug and 1 over square root of n control, OK? That's what the central limit theorem for both x bar drug and x bar control. Two central limit theorems are actually telling. OK. And then we can conclude that this happens.

And as you said, we're trying to make it a bit harder to conclude this. Because let's face it. If we were actually using two simple heuristic, right? For simplicity, right? So I can rewrite x bar drug less than x bar control minus this something that goes to 0. I can write it as x bar drug minus x bar control less than something negative, OK? This little something, OK?

So now let's look at those guys. This is the difference of two random variables. From the central limit theorem, they should be approximately Gaussian each. And actually, we're going to think of them as being independent. There's no reason why the people in the control group should have any effect on what's happening to the people in the test group. Those people probably don't even know each other. And so, when I look at this, this should look like n 0 with some mean and some variants, let's say I don't know what it is, OK?

The mean I actually know. It's mu drug minus mu control, OK? So if they were to plot the PDF of this guy, it would look like this. I would have something which is centered at mu drug minus mu control. And it would look like this, OK?

Now let's say that mu drug is actually equal to mu control. That this pharmaceutical company is a huge scam. And they really are trying to sell bottled corn syrup for $200 a pop, OK? So this is a huge scam. And the true things are actually equal to 0. So this thing is really centered about 0, OK?

Now, if were not to do this, then basically, half of the time I would actually come up with a distribution that's above this value. And half of the time I would have something that's below this value, which would mean that half of the scams would actually go through FDA if I did not do this. So what I'm trying to do is to say, well, OK. You have to be here, so that there is actually a very low probability that just by chance you end up being here. And we'll make all the statements extremely precise later on.

But I think the drug thing makes it interesting to see why you're making it hard, because You don't want to allow people to sell a thing like that. Before we go more into the statistical thinking associated to tests, let's just see how we would do this quantification, right? I mean after all, this is what we probably are the most comfortable with at this point. So let's just try to understand this.

And I'm going to make the statisticians favorite test, which is the thing that obviously you do at home all the time every time you get a new quarter, is testing whether it's a fair coin or not. All right? So this test, of course, exists only in textbooks. And I actually did not write this slide. I was lazy to just replace all this stuff by the Cherry Blossom Run.

So you have a coin. Now you have 80 observations, x1 to x80. So n is equal to 80. I have x1, xn, IID, Bernoulli p. And I want to know if I have a fair coin. So in mathematical language, I want to know if p is equal to 1/2.

Let's say this is just the heads, OK? And a biased coin? Well, maybe you would potentially be interested whether it's biased one direction or the other. But not being a fair coin is already somewhat of a discovery, OK? And so, you just want to know whether p is equal to 1/2 or p is not equal to 1/2, OK?

Now, if I were to apply the very naive first example to not reject this hypothesis. If I run this thing 80 times, I need to see exactly 40 heads and 40 tales. Now this is very unlikely to happen exactly. You're going to have close to 40 heads and close to 40 tails, but how close should those things be? OK? And so, the little something is going to be quantified by exactly this, OK?

So now here, let's say that my experiment gave me 54 heads. That's 54? Yeah. Which means that my xn bar is 54 over 80, which is 0.68. All right? So I have this estimator. Looks pretty large, right? It's much larger than 0.5, so it does look like, and my mom would certainly conclude, that this is a biased coin for sure, because she thinks I'm tricky. All right.

So the question is, can this be due to chance? Can this be due to chance alone? Like what is the likelihood that a fair coin would actually end up being 54 times on heads rather than 40? OK? And so, what we do is we say, OK, I need to understand, what is the distribution of the number of times it comes on heads? And this is going to be a binomial, but it's a little annoying to play with. So we're going to use the central limit theorem that tells me that xn bar minus p divided by square root of p1 minus p is approximately distributed as an n01. And here, since n is equal to 80, I'm pretty safe that this is actually going to work.

And I can actually use [INAUDIBLE], and put xn bar here. [INAUDIBLE] tells me that this is OK to do. All right.

So now I'm actually going to compute this. So here, I know this. This is square root of 80. This is a 0.68. What is this value here? We'll talk about it. Well, we're trying to understand what happens if it is a fair coin, right? So if fair, then p is equal to 0.5, right? So what I want to know is, what is the likelihood that a fair coin would give me 0.68? Let me finish.

All right. What is the likelihood that a fair coin will allow me to do this, so I'm actually allowed to plug-in p to be 0.5 here? Now, your question is, why do I not plug-in p to be 0.5? But you can. All right. I just want to make you plug-in p at one specific point, but you're absolutely right.

OK. Let's forget about your question for one second. So now I'm going to have to look at xn bar minus 0.5 divided by xn bar 1 minus xn bar. Then this thing is approximately Gaussian and 0,1 if the coin is fair. Otherwise, I'm going to have a mean which is not zero here. If the coin is something else, whatever I get here, right? Let's just write it for one second.

Let's do it. So what is the distribution of this if p-- so that's p is equal to 0.5. OK? Now if p is equal to 0.6, then this thing is just, well, I know that this is equal to square root of n xn bar minus 0.6, divided by xn bar 1 minus xn in the bar squared root, plus-- well, now the difference. Is So square root of n, 0.6 minus 0.5, divided by square root of xn bar 1 minus xn bar, right? Now if p is equal to 0.6, then this guy is n 0,1, but this guy is something different. This is just a number that depends on square root of n. It's actually pretty large.

So if I want to use the fact that this guy has a normal distribution, I need to plug-in the true value here. Now, the implicit question that I got was the following. It says, well, if you know what p is, then what's actually true is also this. If p is equal to 0.5, then since I know that root n xn bar minus p divided by square root of p 1 minus p is some n 0, 1, it's also true that square root of n xn bar minus 0.5 divided by square root of 0.5 1 minus 0.5 is n 0,1, right? I know what p is. I'm just going to make it appear.

OK. And so, what's actually nice about this particular [INAUDIBLE] experiment is that I can check if my assumption is valid by checking whether I'm actually-- so what I'm going to do right now is check whether this is likely to be a Gaussian or not, right? And there's two ways I can violate it. By violating mean, but also by violating the variance. And here, what I did in the first case, I said, well I'm not allowing you to check whether you've violated the variance. I'm just plugging whatever variance you're getting. Whereas here, I'm saying, well, there's two ways you can violate it. And I'm just going to factor everything in.

So now I can plug-in this number. So this is 80. This is 0.68. So I can compute all this stuff. I can compute all this stuff here as well. And what I get in this case, if I put the xn bar 1, I get 3.45, OK?

And now I claim that this makes it reasonable to reject the hypothesis that p is equal to 0.5. Can somebody tell me why?

**STUDENT:** It's pretty big.

**PROFESSOR:** Yeah, 3 is pretty big. So it's very unlikely. So this number that I should see should look like the number I would get if I asked a computer to draw one random Gaussian for me. This number, when I draw one random Gaussian, is actually a number with 99.9% this number will be between negative 3 and 3. With 78% it's going to be between negative 2 and 2. 68% is between minus 1 and 1. And with like 90% it's between minus 2 and 2.

So getting a 3.45 when you do this is extremely unlikely to happen, which means that you would have to be extremely unlucky for this to ever happen. Now, it can happen, right? It could be the case that you flip 80 coins and 80 of them are heads. With what probability does this happen? 1 over 2 to the 80, right? Which is probably better off playing the lottery with this kind of odds, right? I mean, this is just not going to happen, but it might happen.

So we cannot remove completely the uncertainty, right? It's still possible that this is due to noise. But we're just trying to make all the cases that are very unlikely go away, OK? And so, now I claim that 3.45 is very unlikely for a Gaussian. So if I were to draw the PDF of a standard Gaussian, right? So n 0, 1, right? So that's PDF of n 0, 1. 3.73 is basically here, OK? So it's just too far in the tails. Understood?

Now I cannot say that the probability that the Gaussian is equal to 373 is small, right? I just cannot say that, because it's 0. And it's also 0 for the probability that it's 0, even though the most likely values are around 0. It's the continuous random variable. Any value you give me, it's going to happen with probability zero.

So what we're going to say is, well, the fluctuations are larger than this number. The probability that I get anything worse than this is actually extremely small, right? Anything worse than this is just like farther than 3.73. And this is going to be what we control. All right? So in this case, I claim that it's quite reasonable to reject the hypothesis.

Is everybody OK with this? Everybody find this shocking? Or everybody has no idea what's going on? Do you have any questions? Yeah?

**STUDENT:** Regarding the case of p, where minus p isn't close to xn. If you use 1 minus p as 0.5, then you're dividing by a larger number than you would if you used xn. So it feels like our true number is not 3.45. It's something a little bit smaller than 3.45 for the distribution to actually be like 1/2. Because it seems like we're adding an unnecessary extra error by using xn bar. And we're adding an error that makes it seem that our result was less likely than it actually was.

**PROFESSOR:** That's correct. And you're right. I didn't want to plug-in the p everywhere, but you should plug it in everywhere you can. That's for sure, OK? So let's agree on that. And that's true that it makes the number a little bigger. You compute how much you would get, we would get if we 0.5 there. Well, I don't know what the square root of 80 is. Can somebody compute quickly? I'm not asking you to do it. But what I want is two times square root of 80 times 0.18. 3.22

OK. I can make the same cartoon picture with 3.22. But you're right. This is definitely more accurate. And I should have done this. I didn't want to get the confused message, OK?

All right. So now here's a second example that you can think of. So now I toss it 30 times. Still in the realm of the central limit theorem. I get 13 heads rather than 15. So I'm actually much closer to being exactly at half. So let's see if this is actually going to give me a plausible value.

So I get 0.33 in average. If the truth was 0.5, I would get something like 0.77. And now I claim that 0.77 is a plausible realization for some standard Gaussian, OK? Now, 0.77 is going to look like it's here. So that could very well be something that just comes because of randomness.

And again, if you think about it. If I told you, you were expecting 15, you saw 13, you're happy to put that on the account of randomness. Now of course, the question is going to be, where do I draw the line? Right? Is 12 the right number? Is 11? Is 10? What is it?

So basically, the answer is it's whatever you want to be. The problem it's hard to think on the scale, right? What does it mean to think on the scale? If I can't think in this scale, I'm going to have to think on the scale of 80 of them. I'm going to have to think on the scale of running 100 coin flips. And so, this scale is a moving target all the time. Every time you have a new problem, you have to have a new skill in mind. And it's very difficult.

The purpose of statistical analysis, and in particular this process that content that takes your x bar and turns it into something that should be standard Gaussian, allows you to map the value of x bar into a scale that is the standard scale of the Gaussian. All right? Now, all you need to have in mind is, what is a large number or an unusually large number for a Gaussian? That's all you need to know.

So here, by the way, 0.77 is not this one, because it was actually negative 0.77. So this one. OK. So I can be on the right or I can be on the left of zero. But they are still plausible. So understand you could actually have in mind all the values that are plausible for a Gaussian and those that are not plausible, and draw the line based on what you think is the right number. So how large should a positive value of a Gaussian to become unreasonable for you? Is it 1? Is it 1.5? Is it 2? Stop me when I get there. Is it 2.5? Is it 3?

**STUDENT:** I think 2.5 is definitely too big.

**PROFESSOR:** What?

**STUDENT:** Doesn't it depend on our prior? Let's say we already have really good evidence at this point [INAUDIBLE]

**PROFESSOR:** Yeah, so this is not Bayesian statistics. So there's no such thing as a prior right now. We'll get there. You'll have your moment during one short chapter. So there's no prior here, right? It's really a matter of whether you think is a Gaussian large or not. It's not a matter of coins. It's not a matter of anything.

Now I've just reduced it to just one question. So forget about everything we just said. And I'm asking you, when do you decide that a number is too large to be reasonably drawn from a Gaussian? And this number is 2 or 1.96. And that's basically the number that you get from this quintel. We've seen the 1.96 before, right? It's actually q alpha over 2, where alpha is equal to 5%. That's a quintel of a Gaussian.

So actually, what we do is we map it again. So are now at the Gaussians. And then we map it again into some probabilities, which is the probability of being farther than this thing. And now probabilities, we can think. Probability is something that quantifies my error. And the question is what percentage of error am I willing to tolerate.

And if I tell you 5%, that's something you can really envision. What it means is that if I were to do this test a million times, 5% of the time I would expose myself to making a mistake. All right. That's all it would say. If you said, well, I don't want to account for 5%, maybe I want 1%, then you have to move from 1.94 to 2.5. And then if you say at I want 0.01%, then you have to move to an even larger number. So it depends.

But stating this number 1%, 5%, 10% is much easier than seeing those numbers 1.96, 2.5, et cetera. So we're just putting everything back on the scale. All right.

To conclude, this, again, as we said, does not suggest that the coin is unfair. Now, it might be that the coin is unfair. We just don't have enough evidence to say that. And that goes back to your question about, why are we siding with the fact that we're making it harder to conclude that the runners were faster? And this is the same thing. We're making it harder to conclude that the coin is biased. Because there is a status quo. And we're trying to see if we have evidence against the status quo. The status quo for the runners is they ran the same speed. The status quo for the coin, we can probably all agree is that the coin is fair.

The status quo for a drug? I mean, again, unless you prove me that you're actually not a scammer is that the status quo is that this is maple syrup. There's nothing in there. Why would you? I mean, if I let you get away with it, you would put corn syrup. It's cheaper. OK.

So now let's move on to math. All right. So when I started doing mathematics, I'm going to have to talk about random variables and statistical models. And here, there is actually a very simple thing, which actually goes back to this picture. A test is really asking me if my parameter is in some region of the parameter set or another region of the parameter set, right? Yes/no.

And so, what I'm going to be given is a sample, x1, xn. I have a model. And again, those can be braces depending on the day. And so, now I'm going to give myself theta 0 and theta 1 to this joint subset. OK. So capital theta here is the space in which my parameter can live.

To make two disjoint subsets, I could just split this guy in half, right? I'm going to say, well, maybe it's this guy and this guy. OK. So this is theta 0. And this is theta 1.

What it means when I split those two guys, in test, I'm actually going to focus only on theta 0 or theta 1. And so, it means that a priori I've already removed all the possibilities of theta being in this region. What does it mean? Go back to the example of runners.

This region here for the Cherry Blossom Run is the set of parameters, where mu was larger than 103.5, right? We removed that. We didn't even consider this possibility. We said either it's less-- sorry. That's mu equal to 103.5. And this was mu less than 103.5, OK?

But these guys were like if it happens, it happens. I'm not making any statement about that case. All right? So now I take those two subsets. And now I'm going to give them two different names, because they're going to have an asymmetric role.

h0 is the null hypothesis. And h1 is the alternative hypothesis. h0 is the status quo. h1 is what is considered typically as scientific discovery.

So if you're a regulator, you're going to push towards h0. If you're a scientist, you're going to push towards h1. If you're a pharmaceutical company, you're going to push towards h1. OK?

And so, depending on whether you want to be conservative-- oh, I can find evidence in a lot of data. As soon as you give me three data points, I'm going to be able to find evidence. That means I'm going to tend to say, oh, it's h1. But if you say you need a lot of data before you can actually move away from the status quo, that's age h0, OK? So think of h0 as being status quo, h1 being some discovery that goes against the status quo. All right?

So if we believe that the truth theta is either in one of those, what we say is we want to test h0 against h1. OK. This is actually wording. So remember, because this is how your questions are going to be formulated. And this is how you want to probably communicate as a statistician. So you're going to say I have the null and I have an alternative. I want to test h0 against h1. I want to test the null hypothesis against the alternative hypothesis, OK?

Now, the two hypotheses I forgot to say are actually this. h0 is that the theta belongs to theta 0. And h1 is that it theta belongs to theta 1. OK. So here, for example, theta was mu. And that was mu equal to 103.5. And this was mu less than 103.5. OK? So typically, they're not going to look like thetas and things like that. They're going to look like very simple things, where you take your usual notation for your usual parameter and you just say in mathematical terms what relationship this should be satisfying, right?

For example, in the drug example, that would be mu drug is equal to mu control. And here, that would be mu drug less than mu control. The number of expectorations for people who take the drug for the cough syrup is less than the number of expectoration of people who take the corn syrup, OK?

So now what we want to do. We've set up our hypothesis testing problem. You're a scientist. You've set up your problem. Now what you're going to do is collect data. And what you're going to try to find on this data is evidence against h0. And the alternative is going to guide you into which direction you should be looking for evidence against this guy. All right?

And so, of course, the narrower the alternative, the easier it is for you, because you just have to look at the one possible candidate, right? But typically, h1 is a big group, like less than. Nobody tells you it's either it's 103.5 and 103. People tell you it's either 103.5 or less than 103.5. OK. And so, what we want to do is to decide whether we reject h0. So we look for evidence against h0 in the data, OK?

So as I said, h0 and h1 do not play a symmetric role. It's very important to know which one you're going to place as h0 and which one you're going to place at h1. If it's a close call, you're always going to side with h0, OK? So you have to be careful about those. You have to keep that in mind that if it's a close call, if data does not carry a lot of evidence, you're going to side with h0. And so, you're actually never saying that h0 is true. You're just saying I did not find evidence against h0. You don't say I accept that h0. You say I failed to reject h0. OK.

And so one of the things that you want to keep in mind when you're doing this is this innocent until proven guilty. So if you come from a country, like America, there's such a thing. And in particular, lack of evidence does not mean that you are not guilty, all right? OJ Simpson was found not guilty. It was not found innocent, OK?

And so, this is basically what happens is like the prosecutor brings their evidence. And then the jury has to decide whether they were convinced that this person was guilty of anything. And the question is, do you have enough evidence? But if you don't have evidence, it's not the burden of the defender to prove that they're innocent. Nobody's proving their innocent. I mean, sometimes it helps. But you just have to make sure that there's not enough evidence against you, OK? And that's basically what it's doing. You're h0 until proven h1.

So how are we going to do this? Well, as I said, the role of estimators in hypothesis testing is played by something called tests. And a test is a statistic. Can somebody remind me what a statistic is? Yep?

**STUDENT:** The measure [INAUDIBLE]

**PROFESSOR:** Yeah, that's actually just one step more. So it's a function of the observations. And we require it to be measurable. And as a rule of thumb, measurable means if I give you data, you can actually compute it, OK? If you don't see a [INAUDIBLE] or an [INAUDIBLE], you don't have to think about it. All right.

And so, what we do is we just have this test. But now I'm actually asking only from this test a yes/no answer, which I can code as 0, 1, right? So as a rule of thumb, you say that, well, the test is equal to 0 then h0. The test is equal to 1 at h1. And as we said, is that if the test is equal to 0, it doesn't mean that a 0 is truth. It means that I feel to rejected h0. And if the test is equal to 1, I reject h0.

So I have two possibilities. I look at my data. I turn it into a yes/no answer. And yes/no answer is really h0 or h1, OK? Which one is the most likely basically. All right.

So in the coin flip example, our test statistic is actually something that takes value 0, 1. And anything, any function that takes value at 0, 1 is an indicator function, OK? So an indicator function is just a function. So there's many ways you can write it. So it's a 1 with a double bar. If you aren't comfortable with this, it's totally OK to write i of something, like i of a. OK. And that's what? So a, here, is a statement, like an inequality, an equality, some mathematical statement, OK? Or not mathematical. I mean, "a" can be, you know, my grandma is 20 years old, OK? And so, this is basically 1 if a is true, and 0 if a is false. That's the way you want to think about it.

This function takes only two values, and that's it. So here's the example that we had. We looked at whether the standardized xn bar, the one that actually is approximately n 0,1 was larger than something in absolute value, either very large or very small, but negative. I'm going back to this picture. We wanted to know if this guy was either to the left of something or to the right of something, right? Was it in these regions?

Now this indicator, I can view this as a function of x bar. What it does, it really splits the possible values of x bar, which is just a real number, right? In two groups. The groups on which they lead to a value, which is 1. And the groups on which they lead to value, which is 0, right?

So what it does is that I can actually think of it as the real line, x bar. And there's basically some values here, where I'm going to get a 1. Maybe I'm going to get a 0 here. Maybe I'm going to get a 0. Maybe I'm going to get a 1. I'm just splitting all possible values of x bar. And I see whether to spit out the side which is 0 or which is 1.

In this case, it's not clear, right? I mean, the function is very nonlinear. It's x bar minus 0.5 divided by the square root of x bar 1 minus x bar. If we put the p in the denominator, that would be clear. That would just be exactly something that looks like this. The function would be like this. It would be 1 if it's smaller than some value. Less than 0 if it's in between two values. And then 1 again. So that's psi, OK?

So this is 1, right? This is 1. And this is 0. So if x bar is too small or if x bar is too large, then I'm getting a value 1. But if it's somewhere in between, I'm getting a value 0. Now, if I have this weird function, it's not clear how this happened.

So the picture here that I get is that I have a weird non-linear function, right? So that's x bar. That's square root of n x bar n 0.5 divided by the square root of x bar n 1 minus x bar n, right? That's this function. A priori, I have no idea what this function looks like. We can probably analyze this function, but let's pretend we don't know. So it's like some crazy stuff like this.

And all I'm asking is whether in absolute value it's larger than c, which means that is this function larger than c or less than minus c? The intervals on which I'm going to say 1 are this guy, this guy, this guy, and this guy. OK. And everywhere else, I'm seeing 0. Everybody agree with this? This is what I'm doing.

Now of course, it's probably easier for you to just package it into this nice thing that's just either larger than c, an absolute value, or less Than C. I want to have to plot this function. In practice, you don't have to.

Now, this is where I am actually claiming. So here, I actually defined to you a test. And I promised, starting this lecture, by saying, oh, now we're going to do something better than computing the averages. Now I'm telling you it's just computing an average. And the thing is the test is not just the specification of this x bar. It's also the specification of this constant c. All right? And the constant c was exactly where our belief about what a large value for a Gaussian is. That's exactly where it came in. So this choice of c is basically a threshold at which we decide above this threshold this isn't likely to come from a Gaussian. Below this threshold we decide that it's likely to come from a Gaussian. So we have to choose what this threshold is based on what we think likely means.

Just a little bit more of those things. So now we're going to have to characterize what makes a good test, right? Well, I'll come back to it in a second. But you could have a test that says reject all the time. And that's going to be bad test, right? The FDA is not implementing a test that says, yes all drugs work, now let's just go to Aruba, OK?

So people are trying to have something that tries to work all the time. Now FDA's not either saying, let's just say that no drugs work, and let's go to Aruba, all right? They're just trying to say the right thing as often as possible. And so, we're going to have to measure this.

So the things that are associated to a test are the rejection region. And if you look at this x in en, such that psi of x is equal to 1, this is exactly this guy that I drew. So here, I summarized the values of the sample into their average. But the values of the sample that I collect will lead to a test that says 1. All right? So this is the rejection region.

If I collect a data point, technically I have-- so I have e to the n, which is a big space like this. So that's e to the n. Think of it as being the space of xn bars. And I have a function that takes only value 0, 1. So I can decompose it into this part where it takes value 0 and the part where it takes value 1. And those can be super complicated, right?

Can have a thing like this. Can have some weird little islands where it takes value 1. I can have some islands where it's takes value 0. I can have some weird stuff going on. But I can always partition it into the value where it takes value 0 and the value where it takes value 1. And the value where it takes 1, if psi is equal to 1, this is called the rejection region of the plot, OK? So just the samples that would lead me to rejecting.

And notice that this is the indicator of the rejection region. The test is the indicator of the rejection region.

So there's two ways you can make an error when there's a test. Either the truth is in h1, and you're saying actually it's h1. Or the truth is in h1, and you say it's h0. And that's how we build-in the asymmetry between h0 and h1. We control only one of the two errors. And we hope for the best for the second one.

So the type 1 error is the one that says, well, if it is actually the status quo, but a claim that there is a discovery-- if it's actually h0, but I claim that I'm in h1, then I admit I commit a type I error. And so the probability of type I error is this function alpha of psi, which is the probability of saying that psi is equal to 1 when theta is in h0.

Now, the problem is that this is not just number, because theta is just like moving all over h0, right? There's many values that theta can be, right? So theta is somewhere here. I erased it, OK.

All right. For simplicity, we're going to think of theta as being mu and 103.5, OK? And so, I know that this is theta 1. And just this point here was theta 0, OK? Agreed? This is with the Cherry Blossom Run.

Now, here in this case, it's actually easy. I need to compute this function alpha of psi, which maps theta in theta 0 to p theta of psi equals 1. So that's the probability that I reject when theta is in h0. Then there's only one of them to compute, because theta can only take this one value. So this is really 103.5. OK. So that's the probability that I reject when the true mean was 103.5.

Now, if I was testing whether-- if h0 was this entire guy here, all the values larger than 103.5, then I would have to compute this function for all possible values of the theta in there. And guess what? The worst case is when it's going to be here. Because it's so close to the alternative that that's where I'm making the most error possible.

And then there's the type 2 error, which is defined basically in the symmetric ways. The function that maps theta to the probability. So that's the probability of type 2 errors. The probability that I fail to reject h0, right? If psi is equal to 0, I fail to reject h0. But that actually came from h1, OK?

So in this example, let's clear. If I'm here, like if the true mean was 100, I'm looking at the probability that the true mean is actually 100, and I'm actually saying it was 103.5. Or it's not less than 103.5. Yeah?

**STUDENT:** I'm just still confused by the notation. When you say that [INAUDIBLE] theta sub 1 arrow r, I'm not sure what that notation means.

**PROFESSOR:** Well, this just means it's a function that maps theta 0 to r. You've seen functions, right? OK. So that's just the way you write. So that means that's a function f that goes from, say, r r, and that maps x to x squared. OK. So here, I'm just saying I don't have to consider all possible values.

I'm only considering the values on theta 0. I put r actually. I could restrict myself to the interval 0, 1, because those are probabilities.

So it's just telling me where my function comes from and where my function goes to. And beta is a function, right? So beta psi of theta is just the probability that theta is equal to 1. And I could define that for all thetas-- sorry. If psi is equal to 0 in this case. And that could define that for all thetas. But the only ones that lead to an error are the thetas that are in h1. I mean, I can define this function. It's just not going to correspond to an error, OK?

And the power of a test is the smallest-- so the power is basically 1 minus an error. 1 minus the probability of an error. So it's the probability of making a correct decision, OK? So it's the probability of making a correct decision under h1, that's what the power is. But again, this could be a function. Because there's many ways that can be in h1 if h1 is an entire set of numbers. For example, all the numbers there are less than 103.5.

And so, what I'm doing here when I define the power of a test, I'm looking at the smallest possible of those values, OK? So I'm looking at this function. Maybe I should actually expand a little more on this. OK.

So beta psi of theta is the probability under theta that psi is equal to 0, right? That's the probability in theta 1, which means then the alternative, that they feel to reject. And I really should, because theta was actually in theta 1, OK?

So this thing here is the probability of type 2 error. Now, this is 1 minus the probability that I did reject and I should have rejected. That's just a little off the complement. Because if psi is not equal to 0, then it's equal to 1. So now if I rearrange this, it tells me that the probability that psi is equal to 1-- this is actually 1 minus beta psi of theta. So that's true for all thetas in theta 1.

And what I'm saying is, well, this is now a good thing, right? This number being large is a good thing. It means I should have rejected, and I rejected. I want this to happen with large probability. And so, what I'm going to look at is the most conservative choice of this number, right? Rather than being super optimistic and say, oh, but indeed if theta was actually equal to zero, then I'm always going to conclude that-- I mean, if mu is equal to 0, everybody runs in 0 seconds, then I with high probability I'm actually going to make no mistake. But really, I should look at the worst possible case, OK? So what I'm looking at is basically the smallest value it can take on theta one is called power of psi. Power of the test psi, OK? So that's the smallest possible value it can take.

All right. So I'm sorry. This is a lot of definitions that you have to sink in. And it's not super pleasant. But that's what testing is. There's a lot of jargon. Those are actually fairly simple things. Just maybe you should get a sheet for yourself. And say, these are the new terms that I learned. What is their test, rejection region? Probably of type I error, probably of type 2 error, and power. Just make sure you know what those guys are. Oh. And null and alternative hypothesis, OK?

And once you know all these things, you know what I'm talking about. You know what I'm referring to. And this is just jargon. But in the end, those are just probabilities. I mean, these a natural quantities. Just for some reason, people have been used to using different terminology.

So just to illustrate. When do I make a typo 1 error? And when do I not make a type 1 error? So I make a type 1 error if h0 is true and I reject h0, right? So the off diagonal blocks are when I make an error. When I'm on the diagonal terms, h1 is true and I reject h0, that's a correct decision. When h0 is true and I fail to reject h0, that's also the correct decision to make.

So I only make errors when I'm in one of the red blocks. And one block is the type 1 error and the other block is the type 2 error. That's all it means, OK? So you just have to know which one we called one. I mean, this was chosen in a pretty ad hoc way.

So to conclude this lecture, let me ask you a few questions. If in a US court, the defendant is found either say, let's just say for the sake of discussion, innocent or guilty. All right? It's really guilty for not guilty, but let's say innocent or guilty. When does the jury make a type 1 error? Yep? And he's guilty? And he's innocent, right?

The status quo, everybody is innocent until proven guilty. So that's our h0 is that the person is innocent. And so, that means that h0 is innocent. And so, we're looking at the probably of type 1 error, so that's when we reject the fact that it's innocent. So conclude that this person is guilty, OK? So type 1 error is when this person is innocent and we conclude it's guilty.

What is the type 2 error? Letting a guilty person go free, which actually according to the constitution, is the better of the two. All right? So what we're going to try to do is to control the first one, and hope for the best for the second one.

How could the jury make sure that they make no type 1 error ever? Always let the guy go free, right? What is the effect on the type 2 error? Yeah, it's the worst possible, right? I mean, basically, for every guy that's guilty, you let them go. That's the worst you can do.

And same thing, right? How can the jury make sure that there's no type 2 error? Always convict. What is the effect on the American budget? What is the effect on the type 1 error? Right. So the effect is that basically the type 1 error is maximized. So there's this trade off between type 1 and type 2 error that's inherent. And that's why we have this sort of multi objective thing. We're trying to minimize two things at the same time.

And I can't find many ad hoc ways, right? So if you've taken any optimization, trying to optimize two things when one is going up while the other one is going down, the only thing you can do is make ad hoc heuristics. Maybe you try to minimize the sum of those two guys. Maybe you try to minimize 1/3 of the first guy plus 2/3 of the second guy. Maybe you try to minimize the first guy plus the square of the second guy. You can think of many ways, but none of them is more justified than the other.

However, for statistical hypothesis testing, there's one that's very well justified, which is just constrain your type 1 error to be the smallest, to be at a level that you deem acceptable. 5%. I want to convict at most 5% of innocent people. That's what I deem reasonable. And based on that, I'm going to try to convict as many people as they can, all right? So that's called the Nieman Pearson paradigm, and we'll talk about it next time. All right. Thank you.