[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: So today let's talk about analyzing randomized experiments, and imagine you have-- we are going to work mostly with some little nod to other possible design about the simplest design of experiment, which is a completely randomized experiment. So someone gave you a sample. Someone took a sample and then flipped a coin and randomly assigned some of the nt units to the treatment group and nc units to the comparison group. And then you have the data.

That's what it is. You know who got the assignment. You have some outcome that is measured in your data set, and you are wondering what to do with it. The nice thing about experiments is we really need to know nothing about econometrics. Everything you have done about statistics, probability and statistics that we have done till today is sufficient to analyze experiments, at least to analyze the simple ones.

It doesn't mean this is necessarily what people do in practice all the time. It turns out that knowing regressions quickens things up a little bit. But it's really not necessary because completely randomized experiments are very simple. So we're going to see three met-- there are three workhorse methods to look at a simple completely randomized experiments, the Fisher exact test, Neyman's approach, which is looking at average treatment effect, computing average treatment effect, and a regression method, which I'm obviously not going to do today since we haven't seen regression.

We'll be very, very trivial when we have seen regression in all of this. So the first question is-- so I should say one thing in particular to start with. Today, for the most part, I'm going to be thinking about-- and this is what both Fisher and Neyman were thinking about. They were thinking about the sample as the population of interest. So they were thinking of-- let's say Fisher was thinking a lot about agricultural experiments.

So imagine a bunch of fields, and then different fields could have gotten different treatments. They were thinking about the effect of the treatment in this particular sample of fields. Same thing with Neyman. He was interested in another question, but for the particular sample that was given to them. Alternatively, we might be interested in an estimate of the effect for some larger population, super population out of which the sample is being drawn.

And in a lot of cases today in economics and social science, that's what we're thinking about. We're thinking about, say, if I do an experiment about a hundred classes, I'm moderately interested in just a hundred classes. I would like to be thinking that I'm saying something about classes in this general environment in general, so the hundred classes that I have are themselves randomly chosen out of a larger population.

So in general, in economics, the way we are thinking about uncertainty is that it's coming from that. There is a larger population. We are drawing some people from that observation. But you can see that another way to think about it is a smaller sample, and the uncertainty comes in that case from who is treated and who is not treated, because it could have been the other way. When we randomized, we could have gotten another assignment.

So for most of today, I'm going to be reasoning in terms of a fixed population, and what is random is not the potential outcome in this population which is fixed, because that's the population, but it's the assignment who happens to have gotten the treatment and the control. Does that make sense? By the way, it becomes relevant again when you're thinking about-- recently with big data, thinking about the problem in this way becomes very pertinent again because if you're Facebook and you're running an experiment in Facebook on your entire client base, well, that's it.

That's the population. You're not really drawing a random sample from a random population. You might have, but you don't necessarily need to. And in this case, it's not that there is no uncertainty. There is still uncertainty. But it's not coming from we've randomly drawn a sub sample. It's coming from the fact that we've treated these guys. We might as well have treated these guys. The result would have been different every time.

So these people, Fisher and Neyman who are really the two people who help us think about experiment, that's the world they are thinking about where the uncertainty doesn't come from drawing a sub sample of the population but from reassigning within the same population. This is where the similarity ends between the two because they were actually pretty bitterly opposed on what was the question of interest when you have an experiment.

You might think it's not that controversial, but they really were fighting bitterly. So Fisher was interested in the sharp null, what he called the sharp null hypothesis, which is the hypothesis that the treatment that we are providing has no effect on anybody whatsoever. OK? He was interested in that. So note that it's very different from the average treatment has no effect on average. It has no effect on everybody whatsoever.

So basically there is no heterogeneity on treatment effect. Everybody's potential outcome treated is the same as their potential outcome control. There is, of course, variation between people's potential outcome, but the treatment is a pure placebo. So let's say if I give sugar to kids for a cough, suppose it has no treatment effect. It actually might have a placebo effect. Imagine it doesn't. If I give sugar to kids before they go to bed, the effect would be zero on average child, so their probability-- they will cough regardless or they will not cough regardless. OK?

So what is very nice about-- what is very convenient about the sharp null is that we solved this magically, the missing data problem, because under the null we exactly know what the potential outcome would have been, because we observe either yi of 1 or yi of 0. What we are willing what we want to test is the null that they are the same, then we know under the null what they would have been so we can recover the entire counterfactual distribution of outcomes under the null from the data that we actually happen to observe.

Which is extremely convenient because once we know that, under the null, we can compute any statistic we please, because we have the entire distribution. We could do the max. We could do the rank. We could do anything. And in particular, something that seems sensible to use is the absolute difference in means by treatment status.

So it's for the particular, for what I observe, the absolute difference in means treatment status is simply the difference in the average observed outcome for the treated group minus the difference in the observed outcome for the control group, and then take the absolute value of that. It's kind of one possible statistic which seems to make a lot of sense. But anything you want you can compute because you have everything. You have the entire counterfactual distribution under the null.

The nice thing is that we can now compute the probability of the randomization distribution, which is any potential way that the observations could have been assigned to the treatment of the control group of the statistics you're interested, for example, our absolute value of the average difference taking a value as large as the actual value given the treatment assigned. So in practice, what do we do?

We take our observation and we say, well, suppose that, in fact, the randomization would have gone differently and we try one round of the randomization where I'll give you immediately an example that is simpler. This is an example of giving people honey for cough, kids honey for cough. So randomized study where children were given honey or nothing. There is also a third treatment that we are going to ignore, where they were given some medication.

But let's now compare honey or nothing. The main outcome if the severity of the cough the night after the assignment. So from one knock off to six, a lot of cough. And in Benson Rubin, in their book on causality and social science and biomedical science use it to illustrate the Fisher's exact test. They assume first to simplify that we have data only for six kids. OK? That we can write all of the assignment.

So first let's see what we observe. So what we observe is for our six kids, whether they were in the treatment group or in the control group. And then we observe the severity of their cough as reported by the parents. So for the treated kids, we observe their potential outcome if treated. So that's [INAUDIBLE] one is three, [INAUDIBLE] two is five, and the other one is zero, so I guess zero was no cough. And then for the last three kids we only see their potential outcome if not treated, so that gives us y zero.

This is some covariates that we also have that we are going to ignore. And this is what we are actually observing. 3, 5, 0, 4, 0, 1. So now we can say, OK, this is the actual thing that we observed. What's the difference between treatment and control? It's the average absolute value of the difference is one, turns out, when we take the difference between 3 plus 5 plus 0 minus 4. That's 8, 4, 1, 1, 6, so 2 divided by 3. Yeah?

STUDENT: [INAUDIBLE] hospital for the treatment.

ESTHER DUFLO: Oh, so this is that. This is that. We are going to ignore it. Because we have random assignment, we are going to focus-- we could do something with it, which we are not going to do for now. We are going to do nothing with covariates, so ignore that. All we are looking at is the severity of the cough before the treatment. You're right.

What we can say instead of being interested in the severity of the cough after the treatment, I could compute the difference and be interested in that as my outcome. And of course, the sharp null would also help me to say what it would have been under any possible assignment. But let's say now we kind of exploit the randomization. We're just comparing what happens after. So the difference is one in the average severity of the cough. OK. Yeah.

STUDENT: I have a quick question. If we did have information on the cost severity before treatment--

ESTHER DUFLO: We do.

STUDENT: Oh, we do. It's xi.

ESTHER DUFLO: This is xi. It's a covariate.

STUDENT: Would it be called difference in difference? It's a small tangent. I'm just trying to--

ESTHER DUFLO: Yes. It's one thing we could do. So for example, if we, instead of being interested in the severity of the cough after, we constructed a new variable called the difference between the severity of the cough after and the severity of the cough before, and then we took the difference between those two as our statistics, that would be a difference in difference. Now we are just interested in that simple difference, so the absolute value. In fact, it should be absolute value of t obs is 1.

So just looking at it intuitively, it doesn't seem that it helps all that much. It seems that the severity of the cough on average is larger for those who got the cough than those who didn't. So the question is, can we reject that maybe it hurts some people or maybe it has helped some people, hurt some people? So we want to reject the possibility that the treatment-- we want to test the hypothesis, see whether we can reject the hypothesis that the honey has no effect on any kid whatsoever.

So since we know that under the null, what would happen, we can fill out the question marks by saying, well, if under the null this is 3, under the null this is 5, under the null this is 0, under this is 4, 0, and 1. OK? We filled it up. So now we have all of the potential treated and control. Very convenient because now we can say, since we don't have that many observation, we can re-randomize. We can run all of the randomization.

For each of the randomization, compute the statistics, which is the absolute value of the difference or the difference, and then take the absolute value. So what we have in this table is all of the potential outcomes. All of the potential assignment. All of the assignment that could have existed. So we 6 and 6, choose two. And the first line is the one that we actually observe. Sorry. The one that we actually observe is this one, the last one, where the first three units are treated and the next three are not treated. So we already know that the difference in cough severity is one.

And then we can now do the same thing. So imagine that it's the first three units that are not treated and the next three that are treated, then the difference would be minus 1. Then we can do 0, 0, 1, 0, 1, 1. Again, compute the difference between the treated and the control in that particular permutation. For people who know bootstrap, this is not a bootstrap because we are randomizing the assignment. We're keeping the observation fixed. This is a permutation. OK?

So we have enough observation that we don't have too many observations so we can compute all of the possible scenarios and say, given what we observe, under the null, this is how the test statistic would have been distributed. I'm giving you all the possibility for the test statistics so you can plot it if you'd like. It's going to have some shape.

We can also compute the p value for the H0, which is simply the probability that the test statistics is equal or larger to this one. We take the absolute value of that so we can count them. So how many in absolute value are greater or equal to one? 16, I think. Yes. 16. Correct. So the p value is 16 divided by 20.8. So are we rejecting H0?

Not at all. I mean, it would be-- it's a very high p value. In other words, what we are seeing is that, well, under the null, most of the difference are actually larger than the one we observe or equal in what we observe. This suggests that this one is not particularly large, and therefore, it is very likely to have happened under the null. Does that make sense?

So now that we have computing power, we could do this type of test even for many, many, many more observations. Even with more observations, we could, in principle, compute the entire-- all of the possible assignments, and for each of the assignments, compute the statistics we're interested.

In this case, the difference between treated and control under each of the assignment, take the absolute value, plot them, and see whether the one we observed tends to be very much to the right or very much to the left. And if it's not, then if it's right smack in the middle here, it would suggest that this is not very-- we certainly can not reject H0. So that's one way to proceed.

If we had really a lot of observation, it starts to be a lot of potential assignments that you have to play with, so you might not be able to do all the permutation. It becomes large. So you don't have to do all of them. You can do a bunch of them. You can do an assignment, a possible assignment by drawing in your sample of n without replacement, like a particular random computer statistics that you're interested for. That assignment, repeat.

Depending on how patient you are, you can do however many you want. The more you do, the more precise your p value is going to be. Of course, the p value, if you do it for all the assignment, the p value is exact. If not, it comes from some null due to the fact that you've only taken possibly a few of the possible assignment or many of the possible assignments, but you control that so you can just reduce the-- you can reduce the standard error of the p value as much as you want to by adding more observations.

And so for the honey study, actually, the first six observations were a bit misleading for honey because, in fact, when you do this simulated test with-- and you have more observations, your p values are-- using all of the observations and doing simulated p value, you actually find p values that are much lower. So using all of the observations and not just the first six, it turns out that the difference between treatment and control was larger than most of the value for the statistics that we get under each of these random assignment of these fake assignment. Does it make sense? So that's Fisher exact test. Yes?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: So when you do bootstrap, you do the opposite. So bootstrap is also a sampling method, but when you do bootstrap, you take the assignment fixed and you sample from the population you have. So actually, you take the entire-- here bootstrapping would be taking a subsamples of the observations we observe, keeping both the outcome that we observe and the treatment constant. So for example, it would be taking-- we have these six observations here.

If I want to bootstrap from six observations, I would sample-- say, if I want to bootstrap sample that has the same size of six, I would sample six times with replacement, both the Wi and the action potential outcome. So I could end up with 3 times the first one and 2 times the second one. So as you can see, it would be pretty disastrous, potentially. So the bootstrap is still a test that works on large sample. The Fisher exact test is what it is, given the sample.

So the Fisher exact test does not rely on anything asymptotic at all. So contrary to the bootstrap where we sample the entire observation, observe outcome plus the treatment together and we put them in a sample and we rerun in that kind of makeup sample, in the randomization inference or Fisher exact test we keep the observation as they are, all of them, and we just reassign the samples.

Which is the motivation of it is that it was randomly assigned to start with, so it could have been another assignment. Nothing prevents you to use Fisher exact test in settings where the original assignment is not randomized but you think it's as good as randomly assigned for some reason. So the Fisher exact test can be used in other contexts as well, although it was originally motivated by this analysis of randomization.

So it is clear the difference between bootstrap and Fisher exact test? So that's one nice thing you could do. That would allow you to test the sharp null. Or in fact, any other sharp hypothesis you're interested in. You could also be interested in testing the hypothesis that the treatment effect is exactly one. Yes?

STUDENT: So can you just quickly explain how rank was calculated?

ESTHER DUFLO: It's the rank of that statistics in the list, I believe. I didn't cover it. It just happens to be in the table. But another statistic you could do is to say-- because here I actually didn't even think about. I didn't get interested in what this measure is, so I don't exactly know what it is because it has a negative number and it's symmetric. So what it is exactly, I don't know.

But it's another-- the rank, another statistic you could compute is-- so one thing you could do is the absolute rank if you wanted, like where does this guy-- where is the largest observation, for example, the largest observation in the treatment. What is it ranked in the overall distribution? Next time, we'll see Kolmogorov-Smirnoff statistic to compare distribution. You could compute it as well for this.

So there is any number of things you could do. But for now, for this lecture, it's enough for you to think about the absolute value of the difference. Not to worry. That's the big choice, anyway. Well, the big choice is the hypothesis you want to test. So we could test the sharp null, which is the treatment effect is exactly zero. We could test the sharp one, where the treatment effect is exactly one.

Then it also gives us the entire counterfactual because we know that if the treatment effect is exactly one, then the potential outcome treated is three, the potential outcome control is two. So we can also compute the statistic under this hypothesis, so we can calculate the treatment-- we can test for the hypothesis that the treatment effect is exactly two. In fact, in this way, if we are patient enough, we can construct exact confidence intervals, which are basically testing a bunch of observations and seeing when we start rejecting the hypothesis that the treatment effect is exactly that number.

So they are not the same. Conceptually quite different from the confidence interval as we discussed in class before, and as we will see in a moment. But they also give you an interval of sorts. So this is Fisher. Neyman was interested in something different. He was interested in the difference between the average potential outcome treated and the average potential outcome of control, which he called the average treatment effect in the sample.

So this is a different object. He argued this was a more interesting object. Fisher disagreed with it. It's actually probably an object we are more used to think about today, what's the average treatment effect. And then he was also interested in constructing confidence interval for the average treatment effect. And in particular, he was interested in testing whether the average treatment effect was zero.

And you can see that the two things are not the same, so it could be that the average treatment effect is zero even if the treatment effect for each particular individual-- for some particular individual is not zero. In fact, it could be that the treatment effect is different from zero for everybody, but the average still is zero. Some people could be hurt and some people could benefit, and yet the average effect is zero. So he was not interested in imposing homogeneity of the treatment effect.

He wanted to leave the possibility that the treatment effect varied but was interested in the average treatment effect. OK? Basically you have to deal with the missing data problem one way or the other. The Fisher way was just to say, well, I'm interested in testing an hypothesis under which I don't have this problem.

Then Neyman was like, I'm not going to try to say anything about any particular unit, but what can I say about the average treatment effect? What can I say about the average difference between potential treated and potential outcome control for the sample? So they both thought they were wrong. But in fact, I think they were both right in the sense that, depending on what you're interested in, you could be interested in both. In fact, people don't use the Fisher exact test enough in my view. I think it's kind of a nice way to get started.

To see, can I reject the hypothesis treatment effect had no effect whatsoever on anybody? If you can not reject it, it's not a very good start, let's say, for anything else you might want to do. So suppose that we have a completely randomized experiment with nt treatment unit and nc control unit what. Would seem to be a reasonable estimate for this object of interest for the average treatment effect? What seems to make sense, given what we saw last time or just first principle.

STUDENT: Sample mean.

ESTHER DUFLO: Yeah, exactly. The difference is sample mean observed for the observed value. Which we can write here as the tau hat as the average treatment effect-- the average for the observed value for people who happen to be treated, the average the sample mean of the sample value for people who happen to be not treated. So that's the notation that we are going to keep. Just the Y bar obs to differentiate from the potential outcome and Y bar obs.

So from what we saw last week, we have reason to think that that's a probably a pretty good estimator. We can actually prove that it's unbiased for the average treatment effect. How do we do that? Well, we replace yi obs by what it is conceptually, and what it is conceptually is Wi times yi of 1. And yi obs, in the subgroup that happens to be treated-- typo. I thought I had dropped all of them. There is a sum that has gone away.

So same thing for the guys who are-- what is the second term here? For the sample of observation, subsample observation that is in the control group, it's the sum of 1 minus what Wi yi is zero and over nt. What I've done is that I've put nt over n here and nt over n here, and so I have an extra n here that I need to keep at the bottom. Why did I do that?

Well, so in this setting where the population-- in this setting, it's kind of a little bit dangerous when someone comes close to you when you teach, but. In this setting where the population is fixed, there is no randomness in the potential outcome. They are fixed. But what is random is the assignment, what they end up getting. If we take the expectation of this tau hat, the only thing that we have to take expectation over is Wi.

And now we know from spatial distribution lecture that E of Wi, the expectation of that is simply the probability of being assigned. So this is nt over n goes here and nc over n goes here, so these guys drop and we are back to the average treatment effect. So the difference between the mean of the treated group and the mean of the control group is an unbiased estimator for the average treatment effect. That might not be a revelation, but it's good to know that it's actually true.

Now what's the standard error of that guy? Because we're working in a small population, it's actually a little bit tedious to compute. So I'm just going to give you the result and talk through the intuition and not go through any of the math of it. If you're very, very interested in it, it's Appendix B of chapter 6 in the *Causal Inference* book by Imbens and Rubin, and it's a long-- it's just very, very simple algebra but many lines of them.

The issue comes from the fact-- the tediousness comes from the fact that we sample without replacement. So every time we sample one unit, we need to think about what would have happened, about what it does to the rest of the observation. But what we obtain is that the variance of this guy, of this tau hat, is simply the sum of the variance of yi of 1 in the sample, the variance of yi of 0 in the sample, and then this third term over here.

So this seems quite logical that the variance of the estimator would be the sum of the variance. What is this third term here? The third term here is the variance of the unit level treatment effect that we also need to add. It's actually coming with a negative sign, so it is at its lowest if the treatment effect is constant. That means the variance of the estimator is the largest for constant treatment effect and is smaller when the treatment effect is not constant.

So that might seem a bit strange in the sense that the less constant the treatment effect, the lower the variance of our estimator. But the reason is that part of the variance of the estimator-- part of the variance that I observe in the data comes from the actual assignment that we want to keep from the estimator, and part of it comes from the fact that there is actually a genuine variance in the treatment effect from unit to unit. So this corrects for that.

So now how do we estimate that number? So the first one is very easy. We have seen it before. We can replace everything by what we actually observe. So the best estimator for the variance of yi of 0 is in the control group, 1 minus nc minus 1, the size of the control group, and then the sum over yi 0 minus the average observed, that to the power of square.

So that's the best estimate as we've seen. Nice estimators for the variance of yi is zero. Same thing for the variance of yi of 1. Now how about the third guy? This covariance term, the variance of the treatment effect. How do we estimate the variance of the treatment-- the unit level variance of the treatment effect. Do you have ideas for that? What would we seem to need to know in order to be able to estimate the variance of the unit level treatment effect?

Well, in general I would wait, but I could wait for a long time because we are not in good shape. We can't estimate that guy, because we don't see yi of 1 and yi of 0. We don't see anything because we are not willing to do the sharp null. We don't see anything. If we observe the treatment observation, we don't have an observation for the-- if we observe a treatment observation, we know the potential outcome treated.

We don't know what it would have been. We don't have anywhere in the data set that tells us anything like that, so we can't compute that guy. So that's a little bit unfortunate, but a simple solution is to say that it's not there. So what Neyman proposes is to ignore the third guy entirely and use as estimator of the sampling variance for tau, just the first two terms. So there are three justifications. This is what we do today.

And there are three justifications for doing that. There are other things you could do, but this is certainly what people do most constantly. There are three justifications for that. The first one is that that's conservative because that term is negative. So if you were able to calculate it, it will reduce the variance. So if you use an estimated variance for your standard that's larger than what it is in reality, you're not going to pretend that you have more precision than you have.

So as an estimate of the variance, using something that is you know larger than the actual variance is a good place to start with. A second justification is that if the treatment effect is constant, it is actually correct. It's exactly correct. In particular, in the world of Fisher where, under the null, the variance of the treatment effect is constant, then that would be-- the Neyman formula would be correct for the average treatment effect.

In economics we often ignore the treatment effect might not be constant, so this becomes correct. And then there is a third argument which is the most subtle, which is it turns out that it's just a small sample problem because it comes from reassigning. It comes from the fact that third term comes from the fact that we are sampling without replacement. That's why there is this third term that comes in.

If we were thinking of estimating an average treatment effect, not just for this sample but for the entire population that this sample is drawn from originally, then tau hat, the difference between the observed value in the sample of the treatment and the control group, is still the best estimate of the average treatment effect in the population. That's very intuitive because, well, what else is there?

That's the best we have. We can prove it, but probably you take me for granted. So it's still a good estimate of the average, and it turns out that the variance-- the last term drops when you calculate the variance of the estimator thinking about it as the estimator of the treatment effect for the entire population. That last term goes away because you don't get this-- you don't get the benefit of the small sample anymore.

So it is actually, if you're interested in what-- not in the Neyman problem, which is what is the average treatment effect in my group. But what's the average treatment effect in the population from which this group has been randomly drawn, then the Neyman estimator happens to be the unbiased estimator of the variance. I haven't proven it. I've just told it. I've just asserted, so don't expect that you should understand exactly why it is true, but it is true.

That's a third reason why it's a perfectly reasonable place to start from. Now, there are alternatives. People have proposed alternatives, but I don't think it's worth it for us to go into it. So all you need to remember in summary is that if you have a completely randomized experiment and you're interested in an estimate of the average treatment effect in that population, or, in fact, in a larger population where the sample is coming from, just take the difference between the average observed in the treatment group, the average observed in the control group.

That's the difference. This is an unbiased estimate of the average treatment effect and an estimate of its variance is the sum of the variance in each of the groups which we estimate traditionally. OK? So now that we are armed with an estimator for the average treatment effect and its variance, we are home free. We can compute anything that we're interested in. In particular, confidence interval and test whatever hypothesis might be interesting, might be good.

So let's start with confidence interval, and recall our prior definition of confidence interval. That's actually a test of my ability to copy from Sarah's note without making mistakes. It's very low because it took me about 25 tries to get that with getting the parentheses in the right place. But we want to find functions-- maybe I did it wrong. I've forgotten it.

Of the random sample, the function A and B such that the probability is that A, the function A of the random sample, the probability that theta, our estimated parameter that was in Sarah's notation, is in between the function A and the function B, that probability is greater than 1 minus alpha. So that's the confidence interval. That was the definition of confidence interval. I recalled it just because it was quick. It went fast when we studied it from Sarah, and it's a very nice concrete example in this case.

So in this case, we want a confidence interval for our friend tau. So we want to find a lower limit and upper limit such that the probability that tau is between this lower limit and that upper limit is greater than 1 minus alpha, with alpha being a number that's given to us but that we are interested in. For example, 0.05 or 0.10. So we want to find this lower and upper bounds.

The only reason why the lower and upper bounds are random in this case, again, because it's a fixed population, is because the assignment could have been different. So again, it's W that generates the randomness in the upper and lower bound. That's the only reason why there is randomness there. What's the confidence interval?

Ah, I forgot the parentheses here. So same one. Now that we are equipped with a variance, we've gone through the calculation. We started. We don't need to do it again. It's going to be two. It's a typo here, plus here. The confidence interval is going to be tau minus some critical value that I'll come to in a minute times the square root of the estimated variance, and then tau plus the critical value and critical value times the square root of the estimated variance.

As we know, with small sample we will take the critical value from the proper table of t distribution. With larger samples, which are often going to be the case in experiments, we actually analyze. We will use the normal approximation and take the critical value from the standard normal table. For example, 1.645 for an alpha of 10% or minus 1.96 for an alpha of 5%. OK.

So that's for confidence interval. Now we have our mean estimated. We can take our difference in sample average to estimate our average treatment effect, compute the corresponding sample variance, sum them up, then take the square root. We can now calculate the confidence interval for that number at whatever confidence value you're interested in. Hypothesis testing. We can start with a Neyman hypothesis, of course, we could use.

Now we have the apparatus to test any old hypothesis you guys might be interested in. But the Neyman hypothesis is one that we should pay particular reverence to, and that's simply the hypothesis that the average treatment effect is zero versus the alternative that it's not zero. We are not making the hypothesis that is larger or smaller, et cetera, where the alternative is the rest. So following our discussion last week, very natural test statistics for this.

This is going to be the difference in the observed outcome treatment minus control divided by the square root of the calculated variance. OK? We know that it follows a t distribution with n minus 1 degrees of freedom, and with n large enough, it can be approximated by a normal distribution. And that gives you the associated p value for a two sided test for the normal approximation. OK? So we compute the t statistics.

We compute the natural test statistic is now going to be the actual observed difference between treatment and control group divided by the-- so the one nice aspect of teaching this lecture right after the hypothesis testing is I hope it serves as a sort of review of special cases and reasonably easy special case of everything when we talk about confidence interval. Make sense? Questions? Concerns?

So here is an example. That's a sneak peak of the problem set. In particular you're going to see all the answers, but of course, we don't really care about the answers. We care about the code. So this is actually a paper that I wrote a few years ago with Rema Hanna, who was then a graduate student and is now a professor at Kennedy School, and Stefan Ryan, who was then an assistant professor here and is now a professor in Texas, Austin?

STUDENT: Yes.

ESTHER DUFLO: In Texas, Austin, and here is what our experiment was. So a big problem. We need to give you one note of background a big problem that schools in India really face is high teacher absenteeism. When I say high I mean high. Like a quarter of the time you show up, there is no teacher, and a further quarter of the time they are here but they are actually not teaching.

So imagine coming to 1431 and your chance to actually have someone in front of you lecturing is about 1 in 2. So you would assume prima facie that might be an issue. And we worked with a small NGO called Salem India and tried to figure out a way of getting teachers to come more. So we are interested in two things, in a way. One is whether that way would be effective and second is whether, if it was effective, it would, in fact, result in higher test scores after about a year of teachers coming more.

And what was our way while being sort of narrow minded economists, we proposed to actually pay people if they showed up to school. The incentive scheme was actually a little bit complicated. But it turned out that basically if you came 10 days or less in a month-- a month is about 22 days. If you came 10 days or less in a month, you pay some floor. 500 rupees for the month. And then for any extra day above that, you would pay an extra 50 rupees.

Such that if someone comes zero days, they will pay 500 rupees. They won't starve. But above some limit they can earn more every time they come in more. How did we measure presence? We gave them cameras. At the time they were paper camera, actually, and they took a photo of themselves and the kids in the morning and in the afternoon. And then we collected the rolls and analyzed them, et cetera.

So the question is whether doing that for a year had an effect on school presence over a year measured by random checks. So we show up randomly during school time and see whether there is a teacher. And then at the end of the training, twice, at the middle of the school year and at the end of the school year, we did a test to know whether the kids actually learned more in treated school and in the control school.

You could treat that as a clustered experiment where each school is a cluster. In fact, that's how we analyzed it in the paper. But let's forget about that and say, imagine that each school is an observation. So first, for presence, of course, that's what it is. It's the average presence, so each school gets an observation. And I'm going to do the same thing for the students where I'm going to say each school is an observation in the sense that I observe average test scores, and that's an observation for this school.

So what do we do? So this is the data we have. So again, we have some pre-treatment variable, which is on the test scores. I'm not going to use that at all. And then post-treatment, so the variable that matters, the variable that I collect post-treatment is the fraction of time that the school is open during the random checks.

So you see it's 58% of the time in the treatment, 80% of the time in the control, then we have some scores for the test, the fraction of kids who actually take the written test, because the test was administered either written or oral. So there was an oral part that everybody took and a written part that only the kids who could write took.

So this is the score on the written part of the test, and this is their score, including-- this is the score on the written part of the test for people who could write and this is the score on the written part of the test, including zero, for people who can not write. It doesn't really matter, the detail. So you can see that the presence seems to be a little bit larger in the treatment and in the control.

The question is, is it something that is just noise or is it something that-- or can I reject the hypothesis that the average treatment effect is zero? These are the relevant numbers in the notations that we use today. So the average in the control group is 58% open in the control. In the treatment group it's 80% open, so the difference is 22 percentage point. So it would indicate that the average treatment effect is 22 percentage point.

If you do economics, you will become very conversant with the fact that 22 percentage point in this case is about 50%, an increase of about 50% over the control mean. That's something that we need to keep doing, which is comparing the average treatment effect with the control mean to see, is it a small increase or a large increase.

Then we can calculate the sample variance in the treatment group and the control group, sum them up. We have the nc and nt, and we can calculate our Neyman variance to be 0.03 to the power of 2. And with that, we can now calculate the confidence interval, which, for open, is 0.15, 0.28 for a 95% confidence interval.

The average treatment effect on fraction open is 22 percentage point, with a confidence interval of 0.15, 0.28, 95% confidence interval of 0.15, 0.28. So can you reject the hypothesis, for example, that the effect is zero? The average effect is zero. Yeah, we can reject the hypothesis that it's zero. It doesn't belong in the confidence interval.

PROFESSOR: By the way, on the problem set that's due on Wednesday, there's a question. I forget which one it is. That's kind of related to the relationship between confidence interval and hypothesis.

ESTHER DUFLO: Yes. So maybe you knew that so maybe I should not assume this knowledge, which we can straight look. If zero doesn't belong in the confidence interval, we can reject the hypothesis that the effect is zero at 95%. But another way to see it more directly is take the average treatment effect divided by the standard error. That's the T statistics. And the T statistic is going to be greater than 1.96.

Then we can look at the effect for, on other things, the kids' test scores. So the confidence interval for the fraction of kids who actually attempt the written test is minus 0.0313. The T statistic is 0.05 divided by 0.04. No. Am I forgetting a 2? Sorry. So that's one point something. So that's lower than 1.96, so we can reject-- we cannot reject the hypothesis that the program did not affect the fraction of kids who attempted the written test.

The T statistics here is a little bit above 2. The T statistic here is exactly 2. So we can reject if we use a alpha of 0.05, that gives us a statistic of 1.96. We can reject the hypothesis that the treatment effect is 0 for the written fraction of the test or for the written fraction of the test after giving a 0 for anybody who didn't try.

Which is potentially a better measure of outcomes. OK. And then we have the confidence interval here. So this is basically this kind of analysis should really be the first thing that I guess the first thing you do when you have a randomized experiment. If you have any pre-variable, pre-data, it's to check that the observation looked similar in treatment and control group.

So you'd have a test of summary statistic for treatment and control group showing that before the samples, in fact, look similar. You want to do something exactly like that. Maybe a Fisher exact test and then that, or maybe just that. That's kind of the very basic thing one might need to do. Any question here?

STUDENT: So one of those summaries-- one of those statistics fails, we fail to reject the hypothesis?

ESTHER DUFLO: Yeah, for the fraction of kids who take the written test. So actually that's convenient for me because then it means that I can-- if that was affected then this would be very hard to interpret because it would mean that the-- oops. It would mean that the test scores I obtain for kids who actually take the written test are selected.

They are only on the population that are unique when it is selected because I only have the test scores for kids who actually do attempt the written test, and that's not a random sample of kids in the class. That's, of course, the best kids in the class. Now if there are more best kids in the class, I could have-- if kids become sufficiently better due to the treatment that they are more likely to take the written test, then the score on the written test, per se, doesn't mean very much anymore because I conflate two things.

I conflate the fact that there might be a direct effect on the test plus the fact that I'm bringing in people from the lower end of the distribution in the treatment group that I don't bring in the control group. So that creates some bias. So that is something that we'll have a chance to talk about later, which is attrition. So basically there is attrition.

In the written only test scores there is attrition because not everybody who is at risk of taking the test actually takes it. This is asking is whether that attrition is actually systematically different in the treatment and in the control group. And it says that the difference is small in the fraction of kids who take the written test and it's also not significant. So it's, in this particular instance, reassuring. Any more question on this example?

So if you do an experiment or analyze an experiment for your project, then I would want you to do some version of Fisher's exact test either for all of the possible permutation or for a subsample, and then something like that would be kind of a minimum that you'd want to do. Now with experiments, if we take data from the real world, we download them from the real world, we use data that already exists, even if we collect it ourselves.

But we haven't designed an experiment, then that's what it is. We have to analyze it the best way possible way we have it. But with experiments we have the freedom to design it to answer the questions we are interested in. And with this freedom comes responsibility because we have to design the experiments such that we won't be disappointed at the end of it, because if we are disappointed at the end of it, not by the result, but by the fact that we find them very difficult to interpret, it's going to be our fault, because we haven't designed the experiment properly.

So there are very many interesting design questions. We'll get back to them later in the semester. If I have a little time, we'll go back just to the three examples that we talked about last time. For now we are equipped to ask a very simple question. Suppose you're interested in designing a simple, completely randomized experiment to test the Neyman hypothesis for a particular thing.

For example, suppose that I'm about-- I haven't done the experiment yet for the camera and I want to know what should be my sample size, how large should my sample be. So intuitively, what are the ingredients that goes into that calculation? Why do I need to even ask this question?

STUDENT: A very large sample size would be costly both in time and in money.

ESTHER DUFLO: Right. So we don't want infinite sample size because we don't have infinite amount of money, so it's unlikely to be possible. So that's on the one side. And on the other side, yeah?

STUDENT: It's too small but we might get a result that's not actually true.

ESTHER DUFLO: If it's too small, the variance can be large of the estimator because there is some n in the denominator. So the larger the sample size, the lower the variance, therefore the tighter the confidence interval will be for our estimated treatment effect, and the more precision we will have to test hypotheses. So that's exactly the trade off that we are facing is how small a sample can I get away with and say something and be able to say something that is going to be meaningful.

So this is called the power calculation. So this is called a power calculation because the way we think about it is we think about it in term of an hypothesis we will want to test at the end of the day, and what's going to be the power of our test. And this is our last review of Sarah's last lecture where we finish by power of the test. So in this case I didn't even attempt to copy it. I just grabbed it.

So for a sample size of n, what we are going to be observing is treatment allocation for our n observations and the observed outcome. And of course, in order to do power calculations, we must say what is the hypothesis that we are interested in testing. For the rest of today, I'm going to assume that we are interested in testing the Neyman hypothesis.

So the Neyman hypothesis is that if H0, then the average treatment effect is zero against the alternative that it's not zero. Little reminder straight from our previous lecture. If we accept H0 and H0 is true, we are good. There is no error. If we reject H0, and in fact H0 is false, again, there is no error. But there are two types of error. One is where H0 is in fact true, but you reject it. This is type I error.

H0 is, in fact, false, but you accept it. You fail to reject it. So the significance level of the test alpha if the probability of type I error and 1 minus alpha is the confidence level, the operating characteristic of the test beta is the probability of type II error and 1 minus beta is the power. So I call beta not the operating characteristic, but the probability to be disappointed.

So it's the probability that at the end of the day you run your test with critical level of alpha and you don't find anything significant. And so 1 minus beta is the power of the test. When you design experiments, that's really what you're thinking about. What's the chance that you don't find anything significant when, in fact, there is an effect. So what ingredients go into the power calculation?

So of course, we need to pick alpha and beta. Well, either we pick alpha and then either we pick beta and we calculate the sample size we need, or, if for some reason someone has given us the sample size and there is no choice, we can tell them, well, for that sample size, this is what the beta is going to be. OK? This is one for one. I'm going to assume that we want a specific beta, we want a specific power, and we will calculate the sample size that we need to reach it.

So alpha, we kind of usually don't have much of a choice because society doesn't want to conclude that some treatment work when, in fact, it doesn't. So we tend to pick alpha low for a high confidence level. Following Fisher, alpha is often 0.05. Fisher has given us the magic number of 5%. So let's say we pick alpha to be 5%.

Now given alpha, you want to pick n such that if the average treatment effect is, in fact, some value tau, the power of the test will be at least one minus beta for some beta that you decide to reach given that the fraction gamma of the units are assigned to the treatment group. So another way to say it is you need to pick nc and nt, but we're going to take gamma as given.

So in addition, you must assume or know something about the variance of the outcome in each treatment arm. We're going to assume that they are the same and it's some number sigma squared. So in summary, for power calculation, we need to know or impose or assume a bunch of things, and then we're going to get n in exchange. So the ingredient that's going to it is alpha.

That's typically 5%. Beta, you have more flexibility there because you're just taking the risk. But people usually go for a power of 80%. Tau, sigma, and gamma. Gamma you can decide for reasons that are going to appear to be obvious in a minute when I show you the formula. Unless there is a strong reason to have more treatment unit and control unit, it is efficient to have the same number as long as you assume that the sigma squared is the same.

So you're going to assume that-- you're going to often pick that 0.5, but if you want to pick 0.7, you can just do a calculation with 0.7. These two here are the sticking point. So this is 0.05, 80%, let's say 0.5. These guys are a bit of a problem, because where are they going to come from? So that's the problem with power calculation is that there is a lot of nice formula, but it really is black magic because there are two ingredients that come into it that we don't know anything about. If we knew something about them, we wouldn't need to do the experiment in the first place.

So the biggest problem is tau. What is tau is what we think the treatment effect is going to be, in reality, our best guess power of the treatment effect. So where could we find tau? So you couldn't do it because you could have an idea because you've run a pilot. For example, you tried the school experiment in three schools and you have an idea of what the effect was there. Of course, it is going to be full of guesswork because it's a small sample.

You could know it from a previous study because you're just replicating someone else's study. So your best guess then is that you're going to find the same effect. But sometimes we just have no idea because it's the first experiment we're running. We didn't really pilot, or the pilot was so low that we don't really trust the results. So in that case, what do we do?

So one thing you have to wonder about is when you're working with a partner, they always overestimate how effective their program is. So they're going to tell you, yes, the effect is going to be of the camera problem. It's going to reduce absence to zero. So then it would be a treatment effect of, what was it? 42 percentage point.

So of course, if we had powered the experiment for a tau of 42 percentage point, we wouldn't have been able to see-- if the result had in fact be 40 percentage point, our sample would have been very small. So the larger tau is, of course, the smaller the sample size you need because the difference is very big that you're trying to detect or you're trying to-- remember, the test is zero.

We are testing that the treatment effect is zero. And in reality, it's not going to be-- in reality, it's not zero. It's here. So the further away tau is from zero, the easier it is going to become to reject zero. So partners tend to overestimate tau, and therefore they tend to underestimate sample size unit. It's kind of always a little bit of a fight. Another thing you could do is to say, well, what's an interesting tau?

An interesting tau is one which, in a sense, an interesting-- is one where if I found that-- if the effect in my sample was in fact tau, it would prompt action by policymakers. So what is this tau, for example? Well, if the treatment is very, very, very cheap, for example, it's sending a bunch of letters, just the cost of the stamp, then even a small effect of sending these mailings, if you find that there is a very small effect, but there is still an effect, well, it's worth doing because it's very cheap.

So if you have a treatment that is very, very, very cheap, you're interested in detecting, and I'll tell you what detecting is. It's not really scientific part. You're interested in detecting. That's a very small effect. What does detecting it? You're interested in rejecting zero when the effect is, in fact, tau. Now you see that this is more about optics or rhetoric than statistics because it's not because you reject zero that the effect is, in fact, tau.

When you give people a point estimate of the average treatment effect, it comes with this confidence interval. That's what you know. The fact that the confidence level includes or does not include zero shouldn't really tell you much about tau hat, but in practice, that's not how policy works. In practice, policymakers will see tau hat, will see whether you can reject zero or not, and if you can reject zero, will operate on the function of tau hat.

Therefore, a useful benchmark is what is the lowest treatment effect that makes it cost effective to scale up. So of course you don't do necessarily exactly all the calculation, but intuitively this is what you do. If the treatment is very, very cheap, you want to have a large sample size to have a lot of power to detect even a small effect, because even a small effect would be very interesting.

If the treatment is super expensive, then you don't need to do a giant experiment that is going to be able to find very precisely a very small effect, because a very small effect of zero is all the same anyway. You wouldn't really act on it. That make sense? So tau is the biggest problem in power calculation, and the rest of it is kind of a window dressing around this pretty big uncertainty. So it is also worth to see how it varies with different taus and stuff like that.

For sigma, you need to get that from prior data with similar outcome. That's usually easier to get because even if someone has not done an experiment, they might have collected data on test scores, for example, if that's going to be your outcome. So you might be able to have a reasonably good idea of what the variability in the test scores are.

So in addition, design matters. I'll come to it in about five minutes after. So whether the experiment is stratified or clustered matters, and I'll explain that more in a minute. That introduces, again, additional uncertainty because it matters to some extent, but how big this extent is depends, again, on things that you typically do not know so you sort of have to guess.

So now for the formula. So of course, in practice, this is the easy part because you don't need to do any of this calculation. There are any number of power calculators online. Of course we'll do it, theta will do it, et cetera. As long as you provide the ingredients, the formula takes place themselves. It is really worth working through the logic, in particular, the practical example from the statistics lessons.

So our test statistics for the hypothesis that the treatment has no effect, remember, is the observed difference in average outcomes divided by the best estimate of the variance, which is approximately the observed difference outcome divided by-- replace the estimated variance by the two sigma squared divided by nt. We are going to reject this if the statistic is large. For example, we are going if alpha for an alpha of 05, we're going to reject it if the absolute value of the T statistic is above 1.96.

And what we need to know for the power calculation is the probability that this occurs. The probability that given that the treatment effect is, in fact, tau and the variance is, in fact, sigma, the absolute value of the test statistics will be larger than 1.96. So what's the probability of that?

So by the central limit theorem, the difference in means minus the true treatment effect scaled by the true standard error of the difference is approximately normally distributed with n of 0, 1. So we can write it here. That guy is approximately normally distributed, and therefore our test statistics is approximately normally distributed with this mean tau divided by square root, the sum of the square root. Mistake here. This should be a sigma squared over nt plus sigma squared over nc and 1. OK?

So the T statistics we know because we are assuming tau and sigma. So given tau and sigma, we know what our test statistics that we are going to do at the end. At the end, when we have everything, we are going to construct the T statistic with the observed values. We're going to get the observed means divided by our best estimate of the variance, and we know that it's going to be approximately distributed normally with that mean tau over the sum of the two variants and 1.

So now that we know that we're kind of home free, the probability that the T statistic is greater than phi of 1 minus alpha over 2 is phi of minus phi minus 1 of 1 minus alpha over 2 plus the tau. That's on one side. And then on the other side, there is the small chance that it's so low that the absolute value is greater than that. That's very unlikely.

You see, the absolute value of the T statistics are greater than 5. It could be either, that it's lower, that the minus is lower or the plus is above. The minus lower is very unlikely. So I wrote it down but we're going to ignore that. So we want the first term, the one above, to be equal to beta. So phi minus 5 minus equal beta requires that 5 minus 1 of beta is equal to what's inside the argument.

All we need is to rearrange now to get the sample size as this formula. So if you've seen these formulas in the past and we are wondering what they were coming from, this is what is coming from. So as long as you know tau sigma squared, gamma, which is the agency arranged here to go to here. I rearranged the nt and nc, which still has this typo into the gammas, and I get the formula. Of course, you don't never need to type that formula in practice. Yeah?

STUDENT: [INAUDIBLE]

ESTHER DUFLO: Oh, there is a typo. The nt here should be nc The typo is not upstairs, actually, so that's fine. So this second term you can ignore anyway, so don't worry about it, and there is no typo in the first term. So you can rearrange nt and nc as nt is gamma times n and nc is 1 minus gamma times n. So that goes into rearranging that term. OK?

And then this is just phi of minus 5 minus minus 1 is equal to beta. So that means I can take 5 minus 1 of that to remove one of the 5 and it's going. So I have 5 minus 1 of beta is equal to what's inside the argument of the phi, which is what it is plus this term rearranged as a function of gamma instead of nt and nc. OK? And then this is just algebra. Rearranging the term because we are interested in n. What puzzles you?

STUDENT: You didn't explain what gamma is in the second one.

ESTHER DUFLO: Gamma is the fraction of the sample that's assigned to the treatment. So tn is gamma times n. And then c is 1 minus gamma times n. The only thing that happened in this term here is I replaced nt by n times gamma and nc by n times 1 minus gamma, and then I move the terms around. So for example, if gamma is 0.5, this is 0.5 and 0.5 square. That's 0.5.

OK? So we can stop here. Again, you can play with the parameters. You're typically going to play with that, and software will typically give you entire curves of how n varies as a function of beta. You can also get how beta varies as a function of n with very similar calculations or anything else that you're interested.