Now let's work out an example that shows how to use the pairwise independent sampling theorem to actually do some sampling and estimation.

So let's remember that our basic theorem says that if we have n independent random variables-- pairwise independent, with the same mean and variance-- and we look at their average, the probability that their average differs from the mean by more than a given tolerance delta is less than or equal to this formula here. Which is the standard deviation over delta squared times 1 over n.

Now we're just going to be plugging into this formula, but I want to state it here for the record to remember that this is the pairwise independent sampling theorem that we're given. Which is, what we've seen in general, allows us to calculate the degree of confidence we can have, the probability that we have given n or the n that we need given how confident we want to be.

So let's go ahead and do the example. And I want to think about what is the possibility of swimming in the Charles. Now, Charles has a coliform count. Coliform are some rather undesirable bacteria that are associated with fecal matter. And we want to know whether it's safe to swim in the Charles. That's a Petri dish showing a kind of sample of bacteria that you might grow, cultured to see what's going on.

The Environmental Protection Agency requires that the average CMD, the coliform microbial density, on the dish is less than 200. And what we want to do is figure out whether, when we do a sample of CMDs around the river, and we get some numbers out, whether, in fact, we will conclude that the average CMD is less than 200.

We need to convince the EPA of that. Now, we're never going to be certain. But what we're going to do is take 32 measurements at random times and locations around the river. And we're going to collect these 32 measurements of CMD. And it's going to turn out that, although a few of them are over 200, the average is well under 200. The average of the 32 samples that we've taken is 180.

And our task now is to convince the Environmental Protection Agency that, on the basis of our data, that the average in the whole river is really less than 200. Even though where a couple of places it was over 100, but on average it was 180, can we convince the EPA that the actual average is less than 200?

And so we're trying to convince them that our estimate based on the sample is within 20 of the actual average. We got 180, so if our estimate is within 20 of the truth, then in fact, the average is less than 200.

Well how are we going to do that? Well, let's look at the parameters in the same pairwise independent sampling theorem and see what we have. So c is the actual average CMD in the river. That's what we don't know. We're

trying to estimate it.

So our samples correspond to a random variable. We're taking a measurement of the CMD at random time and place. And that defines a random variable whose expectation is the unknown city. So we've defined, by our sampling process, a random variable with mean u. In fact, we've done it with 32 variables.

So n samples mean n mutually independent random variables. All with mean equal to the number that I'm trying to estimate. And A n is the average of the n CMD samples. So we have an A32 that we're trying to understand.

So here's the independent sampling theorem formula. And let's see what I know already. I'm going to plug in the knowns. What I know is that n is 32, mu is the unknown, c, that we're trying to estimate. And the delta that matters to us is 20. Because we want to argue that if our average of 180, our measurement of 180, was within 20 of the truth, then in fact, we're under the 200 that the EPA specifies.

So let's plug in our known parameters, 32 for n and 20 for the tolerance. And they plug in here. And that leaves me with the standard deviation, which the formula requires and I have to plug in. And that is a problem, because we don't know what the standard deviation is.

Now, sometimes you can kind of argue that you can figure out what the standard deviation is because you have a theory of what the random distribution is of these measurements. And therefore, you can calculate what its standard deviation should be.

Other times you can actually take a sample of the deviation of your sample and use that as an estimate of the sample of the actual standard deviation. But that's kind of circular, and we're not going to go there.

But another way to do it is to say that if you had some bounds on the maximum possible discrepancy of your measurements, if you had done these kinds of sampling testings of CMDs for a long time and you had never observed two that were more than 50 apart, then what you could argue is that the range of measurements is going to be only 50.

So what we can do is if we say that L is the maximum possible difference that we'll ever measure among samples, that in fact, what you can say is that the worst possible standard deviation when your random variable is ranging over an interval l is l over 2.

And you can check that algebraically, but for now let's just take that as a fact. If you know that your measurements are going to differ by at most l between max and min, the standard deviation can't be more than l over 2. And if we know that l is 50, then I got a number finally to plug in. Because now I can plug in 25 for sigma.

So let's do that. And when I do that, I come out with this calculation that says that the probability that my average, minus c, was greater than 20, that my A32, which we said was 180, was more than 20 away from the truth is less than 0.05.

Or, flipping it around, the probability that my average is within 20 of the truth is greater than 0.095. And so we would like to be able to say, now, that the probability that the unknown c is the 180 that we measured for A32, plus or minus 20, is at least 95%. That seems to be what the theorem told us.

Let's go back. The theorem says that the probability that A32, which we measured to be 180, minus c is less than or equal to 20, is greater than 0.95. So we should go back and tell the EPA that the probability is that c is less than 200 with probability 0.95.

And we'd be pretty tempted to say that. But it's not right. It's technically the wrong thing to say. And why is that? Well, it's an important idea, which is that we're talking about something other than probability here. We're talking about confidence, not probability. And let's explain that a little bit more.

Here's the issue. The number c is a number in the real world. It's an actual physical quantity which is the average CMD in the river. We don't know what it is. But that does not make it a random variable. It is or it isn't within less than 200 or more than 200 and so on.

What's going on is that we have created a probabilistic model of sampling that is designed to have in our probabilistic model this unknown constant. There's nothing probabilistic about the constant. We've introduced the probability by thinking of our random sampling as random variables.

We control the randomness. We can't say that c is random. Our measurements are random. So the right thing we can say is that the possible outcomes of our sampling process can persuasively be modeled as a random variable. So what we can say is that the probability that our sampling process will yield an average that's within 20 of the true average is at least 0.95.

So that's a funny thing to say. What you do is you go tell the EPA that says, look. We don't know what the real average is. But we have a process that gets the right answer 95% of the time to within plus or minus 20. And we measured it. And our process that we right 95% of the time came in with an answer that said it's less than 200. OK?

Now that's the right thing to say. That's the truth. We're making a probabilistic statement about the general properties of our sampling process, and saying, OK. Our sampling process is usually right. The sampling process said less than 200. So we think that's probably right. But we can't say it is right, and we can't even say it's right with any probability. It's just the way that our mostly reliable process yielded an answer.

And that's an important idea to distinguish. So it's our estimate that that's correct with probability 0.95.

Now this is a long thing to say to the EPA. And what we'd like to go back is language that says that we think that the real average, c, is within 20 of 180 is probably within 20 of 180. Because that's what our tests seem to say.

But we're not allowed to talk about the probability that c has some value or other. So instead we summarize the story about how we measured c using a probabilistic process that's right 95% of the time by saying that c is 180 plus or minus 20 at the 95% confidence level. And that is that's a shorthand way of saying we've got this process that we believe in that measured this unknown quantity and told us what it was.

So the moral here that we'll wrap up this little video with is that when you're told that some fact holds at a high confidence level because of some tester, or some random experiment, or some pollster, you have to remember that what that implies is that somebody designed a random experiment to try to get an estimate of reality.

And you can always question whether you believe in that random experiment. It's important to understand that there is some random experiment back there. And you should be wondering about what is it? And do I believe in it? And even more important question to ask is why are you hearing about this particular experiment? How many other experiments were tried and not reported?

The point is that people can perform various kinds of tests at the 95% or higher confidence level. But when the tests don't come up with an interesting result, they don't bother to publish them or announce them. And of course, when they come up with a surprising result which is going to be wrong one out of 20 times, those are the results that they publish and submit and advertise. Because they sound good.

In fact, the major drug company Glaxo SmithKline, after paying $3 billion as a judgment against them in 2012 for suppressing the negative results of clinical trials, just agreed to now make public in February 2013 all of the clinical trials that they perform. So that you're not just learning about the cherry picked positive results, but about the negative ones as well.

And in fact, [AUDIO OUT[to get this point home, you might want to look at the cartoon at XKCD, which explains how it is that when there's a problem with green jelly beans at the 95% confidence level.