PROFESSOR: Now you may remember a discussion of the birthday paradox, which says that if you have a group of 27 random people. The probability is almost 2/3 that two of them are going to have a matching birthday, even though there are 365 birthdays in the year. You might sloppily think that with 27 people there'd only be a 27 out of 365, or some chance like that. It's actually 2/3.

And by the time you get to a class of 110-- which is what we have data for and we're going to be looking at-- it turns out that the odds are almost 3/4 of a million to one that you'll have a couple of people with matching birthdays. So let's look at the matching birthday problem a little bit more today. And the reason we're looking at it is because it's a lovely example where there really is pairwise independence, and not mutual independence. So it's reinforcing the key idea behind the additivity of variance, and the pairwise independent sampling theorem. We're not going to use the sampling theorem here, but just pairwise independence, but it's worth looking at.

Now before I go further let me mention that the birthday problem is just what we're doing for fun. But in fact, it has some real applications in more than one area, but the most famous one is the so-called birthday attack on a cryptosystem, which involves being able to search for matching pairs of keys with a relatively small sample. And you're very likely to find at least two that match. So with that motivation claimed, but not examined, let's just go back to thinking about birthdays.

OK so let's suppose that I have some group of n people, and there are d-days in the year, just to keep the parameters abstract and not get too stuck on the numbers. Keeping the parameters makes it actually clearer to reason about. So we're implicitly assuming here that each person is kind of a random variable, or a random choice of a birthday. So each of these people are really random variables that return the value of a birthday.

And it is a matter of fact, we're going assume that all the birthdays are equally likely. Real birthdays aren't. They tend to be of January tends to be a popular month, November tends to be a more popular month than other times. But let's ignore that because if the applications in crypto things really are uniform. And it makes our analysis plausible, still plausible but easy if we assume that birthdays are equally likely, OK.

P is the number of pairs of birthdays that match in this population of n people. OK, let's get a

grip on p by thinking of it as a sum of indicator variables. So let's let M sub ij be the indicator variable that ith and jith people among the n have a matching birthday. Well the number of matching birthdays is then simply the sum over all the possible pairs of people of whether or not they have a matching birthday. It's the sum of these indicator variables M sub ij. And the number of these indicator variables is of course all the ways of choosing two out of n people.

So in short, if I look at the expectation M sub ij, let's think about that for a minute. We're assuming that all the birthdays are equally likely. And so I'm asking whether the ith and the jith people have the same birthday. Well whatever the ith's person birthday turns out to be, let's say it's November 5, the jth person, who has a uniform probability of equalling any birthday, still has a uniform probability 1 chance in d of equalling November 5, which happens to be my birthday.

OK so in short the probability that any two people have a matching birthday is one chance in d. And that means that the expectation of the indicator variable for that event, M sub ij, is 1 over d. And that tells us, by linearity of expectation, that the expected number of pairs is simply the number of those pairs times the expected number per pair, and choose 2 times 1 over d.

Well as I said we have data for 110 students. So the expected number of pairs in a collection in a student body of 110 is 110 choose, 2 times 1 over 365, or about 16.4 pairs is the expected number of pairs of matching birthdays. OK, now that's an expected value. How likely is it to be if I take a selection of 110 students, and I count how many pairs of birthdays are there, do I really expect to get close to 16.4 or not?

Well what we're asking for is the probability that p is near its mean, that the distance between P and 16.4 is greater than k. I hope that as k gets bigger this probability is small. And so I'm really quite likely to have close to 16.4 birthdays in my sample of 110. But this probability is one that's a mess to calculate. But we can get a grip on it because the variance of P is easy to calculate. And that will allow us to apply the Chebyshev bound, and get some kind of an estimate on the likelihood that P is near its expectation.

So the key observation that we need is that the indicator variables are pairwise independent. So let's think about the indicator variable for the event that the ith and the jth people have the same birthday, let's call them Albert and Drew. So Albert's the ith person, Drew is the jth person. And I'm interested in the event that Albert and Drew have the same birthday. And let's compare that to another pair of people, and whether or not they have the same birthday.

So let's first of all think about Dave and Mike, whether Dave and Mike have the same birthday. And I want to know if these two events are independent. Well remember we are assuming that Albert's birthday is independent of Drew's birthday is independent of David's, is independent of Mike. Each of the people is supposedly chosen independently, and their birthdays are independent. So it's obvious that these two pairs that don't overlap have nothing to do with each other, and we don't have to worry about that. You could prove that formally, but it is obvious because each of the individual birthdays are independent.

Now what's more interesting is the case when I asked whether or not Albert and Drew having the same birthday is independent of Albert and Mike having the same birthday. And that one is not so obvious. Here's a way to think about what could go wrong. Suppose that in fact the birthdays weren't uniform, suppose that some birthdays were more common than others.

OK that makes it more likely that if Albert and Drew have the same birthday it slants things, so that they're more likely to have this very common birthday than they would have been otherwise. And now once I know that they match, and therefore are more likely to have the common birthday than they would have without any information, I know that Albert is more likely to have this common birthday than otherwise. And that means that Mike is even more likely to match Albert, because Albert's got the common birthday than Mike was to match Albert without any further information about what Albert's likely birthday was. You can think about that, and it can be worked out numerically, easily enough.

So uniform is going to be a crucial factor here in order to conclude that Albert and Drew, and Albert and Mike are mutually independent events. But let's go back and think about it. All we really need is that Mike is uniform in order to conclude that these two events are independent. Because we know that Mike and Andrew and Albert separately are independent of each other. Their birthdays are chosen independently.

So that intuitively means that the probability that Mike has any given birthday doesn't really matter what's going on with Albert and Drew, because Mike is independent of Albert and Drew. And if we know that Mike's probability of having a birthday is uniform, then whatever the birthday that Albert has, whether he matches Drew or not, Mike has a 1 chance in d of hitting the same birthday of whatever Albert wound up having. And that means that the probability that Mike matches Albert is the same one over d than it would have been if we had no further information.

This is an argument that, in fact, is made rigorous in some class problems and a problem set, but let's just take it as plausible enough based on this hand-waving argument that I articulated, that these two events are independent pairwise and so the corresponding indicator variables and M Albert Drew, and M Albert Mike are independent of each other. So that's what we've argued.

But notice that these events of pairwise matching are certainly not three-way independent, because after all if I know that Albert and Drew have the same birthday, and that Albert and Mike have the same birthday, I absolutely know with certainty that Drew and Mike have the same birthday. So this is a very nice, basic example where you have pairwise independence, but not three-way independence, assuming that all of these random variables Albert, Drew, and Mike are uniform in what birthday they have.

OK so let's go back to counting birthdays. The variance of an indicator is pq. So in this case p is 1 over 365, and q is 1 minus 1 over 365. And because of pairwise independence, the variance of p, which is the sum of the M sub ijs, the variance of the number of birthday pairs, is the sum of those variances. It's 110 choose to times the variance of the M sub ij turns out to be about 16.37, which means that the standard deviation sigma is less than 4.

Now I can apply Chebyshev, because by the Chebyshev band the probability that 16.4 is within a 2 sigma, is further away than 2 sigma, is only one chance in four. Which means the probability that it's within 2 sigma, that the actual number of measured pairs is within 2 sigma of the expected number 16.4 is greater than 1 minus 1/4, or 3/4. There's a 3/4 chance that the number of pairs that we find is within 2 sigma of the expected number 16.4. Sigma was about 4, so this is 8, which means that we're expecting, with 3/4 probability, somewhere between 8.4, meaning 9, and 24.4, meaning 25 pairs.

So 75% of the time, in a class of 110, we're going to find between 9 and 25 pairs of birthdays. Did that actually happen? Well it did.

In our class of 110 for whom we had data, we actually found 21 pairs of matching birthdays. Literally we found 12 pairs and three triples, but each triple counts as three matching pairs. And there they are, the blues are triples. And you can see whether your birthday is among those, and knowing that you have a classmate or two that have the same birthday that you do.

So there are 15 different birthdays, but they count as 21 pairs because it's 12 single pairs, and three triplets, each of which counts for three pairs.