In this segment we look into the probability that the sum of n independent identically distributed random variables takes an abnormally large value.

We will get an upper bound on this quantity, which is known as Hoeffding's inequality.

This is an upper bound that applies to a special case, although the method actually generalizes.

Here is the special case that we will consider.

The random variables, the Xi's, are equally likely to take the values minus 1 and plus 1, with equal probability.

And we're interested in the random variable, which is the sum of the X's.

What do we know about this random variable?

Well, the expected value of each one of the Xi's is equal to 0, because the distribution is symmetric.

And also, the distance of Xi from the mean has always magnitude 1.

And for this reason, the variance of the Xi's is equal to 1.

For this reason, the random variable Y has a mean of 0 and a variance equal to n.

Now, what do we know about the random variable Y?

By the central limit theorem, Y has an approximately normal distribution.

The distribution is centered at 0.

And also, the random variable Y over square root of n, this is a standardized random variable.

So it's approximately normal.

And so the probability that this number is larger than or equal to some a is approximately 1 minus the cumulative of the standard normal.

So Phi here stands for the standard normal CDF.

What does this tell us?

It tells us that if we take somewhere here, the number square root of n times a, then this probability down here in the tail is approximately constant, no mater what n is.

And this in particular tells us that values of the order of square root n are fairly likely to occur.

However, what we're interested in here is not being larger than square root n times a.

We're interested in being larger than n times a.

So we're talking about what happens further down in the tail of the distribution.

So if we take here n times a, we're looking at this probability down here.

And we want to ask, how small is that probability?

Well, we have Chebyshev's inequality.

And Chebyshev's inequality tells us that the probability of Y being larger than a certain number is less than or equal to the variance of Y divided by the square of that number.

And in this case, since the variance is n.

This is 1 over n a squared.

So Chebyshev's inequality tells us that this probability goes to 0, and it goes to 0 at least as fast as 1/n goes to 0.

However, it turns out that this is extremely conservative.

Hoeffding's inequality, which we're going to establish, tells us something much stronger.

It tells us that this tail probability down here falls exponentially with n.

So this is what we want to show.

And let us get started to see how the derivation goes.

The derivation relies on a beautiful trick.

Instead of looking at this event here, we're going to look at the following equivalent event.

Let us fix some number s.

We're going to leave the choice of s free for now.

It is only that we're going to assume that s is a positive number.

And throughout, we're also assuming that a is also a positive number.

Now, we look at this quantity, and the event that this quantity is larger than or equal to e to the sn times a.

Now, this sum is larger than or equal to na if and only if this quantity is larger than or equal to e to the sna.

This is because a and s are both positive.

So the direction of the inequalities does not get reversed, and also because the exponential function is monotonic.

So this event is the same as that event.

So we will try to say something about the probability of this event.

How are we going to do it?

We will use the Markov inequality, where Z is the random variable that appears here.

So by Markov's inequality, this probability is less than or equal to the expected value of the random variable that we are dealing with divided by this value.

Now, the exponential of a sum, we can factor it as a product of exponentials.

And then we use the assumption that the X's are independent.

Since the X's are independent, e to the sX1 is independent from e to the sX2 and so on.

And so we have the expected value of a product of independent random variables.

And so this is equal to the product of the expectations.

So we're going to multiply the expected value of e to the sX1 with the expected value of e to the sX2 and so on.

But because all the Xi's are identically distributed, the terms we get are all the same.

So we get this term to the nth power divided, again, by e to the sna.

Or we can write this in more suggestive form, as follows.

It's the expected value of e to the sX1 divided by e to the sa, all of that to the nth power.

So think of that as being some number rho to the nth power.

When is this bound going to be interesting?

It's going to be interesting if rho is less than 1, because in that case, this bound falls exponentially with n.

And so this probability in particular will fall exponentially with n.

The key here is that we have freedom to choose s.

For any value of s, we obtain an upper bound.

We're going to choose s so that we get the most informative or a most powerful upper bound.

So let us continue to see what we can do.

First, let us write down what this expected value is.

Since X1 takes values minus 1 or plus 1 with equal probability, this expectation is the following.

With probability 1/2, X1 takes the value of 1.

And so this random variable is e to the s.

And with probability 1/2, it takes the value minus one, in which case this random variable is e to the minus s.

So this is the expectation in the numerator.

And we write again the term in the denominator.

And we have all this to the power n.

If we can choose s so that this quantity is less than 1, we will have achieved our objective.

Can we do that?

Let's see.

Let's look at the numerator as a function of s.

When s is equal to 0, we have 1 plus 1 divided by 1/2.

That gives us 1.

And then as s moves away from 0, this function will have this kind of shape.

And it is symmetric around 0, because we have an s and a minus s here.

In particular, the derivative of this function is 0 at 0.

Let's look at the denominator term.

The denominator term is an exponential.

a is a positive number, so it's an exponential that has a shape of this kind.

The important thing to notice is that this exponential has a positive derivative at 0.

What does that tell us?

That at least in the vicinity of 0, this term, the denominator, is going to be larger than the numerator term.

And that implies that in the vicinity of 0, this fraction is going to be less than 1.

And we will have achieved our goal of an exponentially decaying bound.

So the conclusion is that for small s, we have that rho is less than 1.

Now, we would like to get an explicit value for rho.

And we will do that by fixing a specific value for s.

It turns out that if we set s to be equal to a, then the bound that we get is going to be that this probability here is less than or equal to e to the minus na squared over 2.

And this is the Hoeffding bound.

At this point, you may just pause.

Or if you're curious, you can continue with this video to see the algebraic manipulations involved in order to show that this expression is less than or equal to that expression.

But before going there, I would like to make a general comment.

Even if the X's had a different distribution but with 0 mean, the derivation up to this point would go through, here you would have a somewhat different expression for the expected value of e to the sX1.

However, it turns out that the expression that you get here will always have this property that it has a 0 derivative.

This is a consequence of the assumption that we assumed zero mean.

And because of that, we will still have a picture of this kind.

And so this fraction will always be less than 1 when we choose s to be suitably small.

And so this is going to give us a result for more general distributions.

And that more general results is known as the Chernoff bound.

However, we will not develop in this video the Chernoff bound in its greater generality.

We will just stay with Hoeffding's inequality that gives us the basic idea.

And what we will do next will be to derive this inequality.

So I'm carrying over what we figured out in the previous slide-- and this is the quantity here that we wish to bound.

We will look at the numerator term.

And we're going to use a Taylor series for the exponential function.

Remember, the Taylor series for the exponential function takes this form.

And using that, we have 1/2 e to be s plus e to the minus s is equal to the following.

We first write the Taylor series for e to the s.

I'm just copying from here.

It's 1 plus s plus s squared over 2 factorial plus s cubed over 3 factorial.

And we continue similarly.

And then for the term e to the minus s, we have a similar expansion, except that we put a minus s in the place of s.

Now, minus s squared is the same as s squared, with a plus sign.

But for s cubed, when we have minus s, this becomes minus s cube and so on.

And so we see that in this expansion here, we will alternate between positive and negative signs.

This means that all of the odd power terms will cancel each other.

But the even power terms will survive.

So what we obtain is the sum of all of those terms.

But we only have the even power terms.

So we have powers of the form 2i.

These are the even integers.

And in the denominators, we will always have the factorial of whatever exponent we have at the top.

Now, let us get a bound on this term in the denominator.

2i factorial is 1 times 2 times 3, all the way up to i.

And then we continue-- i plus 1, i plus 2, all the way up to 2i.

And what we have is, first, i factorial.

But then each one of these terms is larger than or equal to 2.

And we have i such terms.

And this gives us this inequality.

So we're going to use the substitution here.

Because this term is in the denominator, the direction of the inequality is going to be reversed.

And we obtain this.

Now, we can rewrite this by taking this term 2 to the i and combining it with the other term in the numerator.

And what we have is s squared divided by 2-- all of that to the i'th power.

Now, does this expression look familiar?

It is of exactly the same form as this expansion.

But instead of s, we now have s squared over 2.

Therefore, this is equal to e to the s squared over 2.

So we managed to bound this term.

Using now this bound, we go back to this inequality.

And we have that this is less than or equal to-- in the numerator, we have e to the s squared over 2.

In the denominator, we have e to the sa, and all that is raised to the n'th power.

Or another way to write this is, e to the s squared over 2 minus sa, and all that to the n'th power.

And now, if I choose s equal to a, what I obtain here is going to be e to the a over 2 minus a squared.

That leaves me with e to the minus a squared over 2.

And then I take this factor of n as well.

And the final conclusion is that this quantity becomes equal to this term.

And so we have completed the derivation that this expression is less than or equal to this quantity when we choose s equal to a.

And this is Hoeffding's inequality.