Mathematically speaking, the Chebyshev inequality is just a simple application of the Markov inequality.

However, it contains a somewhat different message.

Consider a random variable that has a certain mean and variance.

What the Chebyshev inequality says is that if the variance is small, then the random variable is unlikely to fall too far off from the mean.

If the variance is small, we have little randomness.

And so X cannot be too far from the mean.

In more precise terms, we have the following inequality.

The probability that the distance from the mean is larger than or equal to a certain number is, at most, the variance divided by the square of that number.

So if the variance is small, the probability of falling far from the mean is also going to be small.

And if the number c is large, so that we're talking about a large distance from the mean, then the probability of this event happening falls off at a rate at least 1 over c squared.

By the way, I should add here that c is assumed to be a positive number.

If c was negative, then the probability that we're looking at would be equal to 1 anyway.

And there isn't any point in trying obtain a bound for it.

To prove the Chebyshev inequality, we will apply the Markov equality as follows.

The probability of interest is the same as the probability that the square of this quantity is larger than or equal to the square of c.

But now, here we have a non-negative random variable.

And we can apply the Markov inequality with X replaced by this random variable and with a replaced by c squared.

So this gives us the expected value of the random variable of interest divided by c squared.

But we recognize that the numerator is just the variance.

And this is the Chebyshev inequality that we claimed.

As an application of the Chebyshev inequality, let us look at the probability of this event that the distance from the mean is at least k standard deviations, where k is some positive number.

Using the Chebyshev inequality with c replaced by k times sigma, we obtain sigma squared over c squared, which in our case is k squared times sigma squared, which is 1 over k squared.

So what this is saying is that if you take, for example, k equal to 3, the probability that you fall three standard deviations away from the mean or more, that probability is going to be less than or equal to 1 over 9.

And this is true no matter what kind of distribution you have.

Let us now revisit our earlier example, where X is an exponential random variable.

And we're interested in the probability that the random variable takes a value larger than or equal to a.

The Markov inequality gave us a bound of 1 over a.

And as we recall, the exact answer to this probability was e to the minus a.

Let us see what we can get using the Chebyshev inequality.

Now, our random variable has a mean of 1.

Let us assume that a is bigger than 1, so that we're considering an event that we fall far away from the mean by a distance of at least a minus 1.

That is we write the probability that X is larger than or equal to a as the probability that the distance of X from the mean is larger than or equal to a minus 1.

And now, this event is smaller than the event that the absolute value of X minus 1 is larger than a minus 1.

This is because if this event is true, then that event will also be true.

And now, we can apply the Chebyshev inequality.

Here we have the distance of X from the mean.

So the Chebyshev inequality applied to the random variable X will have up here the variance of X, which is equal to 1.

And in the denominator, we will have a minus 1 squared.

Notice that if a is a large number, this quantity here behaves like 1 over a squared, which falls off much faster than 1 over a.

So at least for large a's, the Chebyshev bound is going to give us a smaller bound and, therefore, more informative than what we obtained from the Markov inequality.

In most cases, the Chebyshev inequality is, indeed, stronger and more informative than the Markov inequality.

And one of the reasons is that it exploits more information about the distribution of the random variable X. That is it uses knowledge, not just about the mean of the random variable.

But it also uses some information about the variance of the random variable.