Module 3: Correlation

Video Transcript

Video 3.1: Module Introduction: Correlation (4:52)

This module is focused on the concept of correlation between random variables. This is the way we use to capture random quantities. Specifically, we will go beyond thinking about one probability distribution or one random variable or one random quantity, and now we will consider two or more random variables and how their respective distributions relate to each other. Or in other words, how two or more distributions or random quantities co-behave. Let's start with a motivating example of why co-behavior of random quantities is so important. Suppose you are exposed to 10 events. Each could cause an injury with a probability of 1/10. And what you are interested in is to assess the aggregated risk, namely the number of injuries that could happen across all of these 10 events. Now the average number of injuries is one, but this does not really capture the aggregated risk. And you would like to understand the distribution of the random number of injuries that you are likely to see. Now, this is highly dependent on what you are going to assume about the co-behavior of the likelihood across all of these 10 events. One natural assumption could be to assume that the likelihood of an injury per event is independent of each other and in which case we are falling back to something that we've already seen, because the way to think about this situation is as a binomial distribution, we already talked about that, that has 10 experiments.

Each could fail or each could succeed depending on how you want to model that with the probability 1/10, which means that each event has a probability of 1/10 to have an injury. In which case as you can see here, in all likelihood or most likely you're going to have one injury, but you could also have zero injuries, you could have three injuries, four injuries. In fact, any number of injuries has a positive probability. And what you can see here is the probability distribution, the histogram that describes the different likelihoods of different numbers. But it could be that you cannot really assume that all of these events are independent of each other or that the injury, the likelihood of an injury per events are independent of each other. Here's another assumption that in some cases could be very relevant and this is a case where we later going to define as these random variables are entirely negatively linearly correlated or perfectly negatively linearly correlated. What do we mean by that? In this case, we assume that we are ensured to have exactly one injury but in what event that's completely random. So, in some sense, we are assuming here, there's going to be exactly one event with probability 1/10 in one of the 10 events that we are going to have. So, in this case, indeed the likelihood of an injury pair event is 1/10. But when you look on the aggregated number of injuries that you're going to have, it's guaranteed to be exactly one with probability one, because you're going to have exactly one event with an injury.

As you can see, this picture or this situation or this scenario is an entirely different than the previous scenario, although the marginal likelihood of an injury pair event is exactly the same 1/10. But this is not the only possible scenario. Here's a third scenario. Again, this is a scenario that we will later define as entirely or perfectly positively correlated events. In which essentially, you are assuming that you will either have no injury in none of the events with probability 90% and with probability 1/10, 10%. You're going to have injuries in all the events. And what is the resulting histogram here. The resulting histogram here is very by-model, you either you have 10 injuries with probability 1/10 or you have zero injuries with probability 90%. And again, as you can appreciate hopefully, this is an entirely different scenario from the perspective of aggregated risk compared to the previous two ones. So hopefully, that drives home the message or the idea that it's very, very important in many cases to understand the co-behavior of random quantities and not just merely talk about the average or even the marginal distributions, the distribution of each one of these random quantities separately. More to come next.

Video 3.2: Portfolio Selection Problem Introduction (8:29)

So, next we are going to discuss a much more positive and optimistic scenario, where you want to invest money in different stocks and you're thinking about the risk now in the context of an upside scenario where you're actually interested in making some money. And specifically, the use case we'll consider three retail sector stocks of Amazon, Home Depot and Walmart. And what you're interested in is to invest money and buy the stocks today and perhaps sell them within one month from today. So, what we would like to see is how the concepts that we talked about so far with additional concepts, like correlation and more generally, how analytics can help inform this kind of decision? So, what is the data? Well, the data that we have here is about 60 months of historical data of the average return or aggregated return over that month for each one of these stocks. As you can see here, each one of the rows represent a month. And for each month you have three numbers that correspond to the return of each one of these stocks: Amazon, Home Depot and Walmart on that month. And so, this is the data that we're going to work with. It's an empirical data.

It is an historical data. So, we are going to work with the induced empirical distributions that this data allows us to create. And what you see here is a visualization of the time series. What do you mean by time series? Since we are looking on temporal data, month after month, after month, sometimes we call this kind of data as a time series of each one of the stocks and the return of each one of the stocks over time. And what you also see here are different statistics of that return. So, the first column is the mean return over the 16 months that we have in the data. We also define the min return and the max return. This is the second and the second to last columns, that basically give us the range, which is the last column, which is the minimum to maximum return values that we've seen throughout the data. And as well, we calculated here all the typical quantities that we've seen before like the 10th quantile, the 25th quantile, the 50th quantile, the 75th quantile and 90th quantile. So, these are typical statistics that we've seen before that we use to describe probability distributions or random variables. So, one way to think about the setting is that each, the return of each one of these stocks is a random variable, and what you see here are statistics of this random variable. But the question is, which statistics are going to be relevant or important to inform the decision that we are interested in this use case, which is investing money with the hope to get the good return on our money one month later?

So, there are many choices here. As we talked about before, you can make different choices about that. But one natural choice is clearly to worry about the expected return, the mean return, which is calculated here for the three stocks: Amazon 3.23%, Home Depot 2.04% and Walmart 1.08%. What is more interesting perhaps is how to think about a statistic that would capture risk. And here we have many, many choices. We are going to focus today in this discussion, about the standard deviation of the random return of each one of the stocks. And what you see here, this is going to be our way to capture risk by calculating the standard deviation of each one of the stocks, 8.40% for Amazon, 5.26% for Home Depot and 5.38% for Walmart. Now, remember that in prior modules, we also discussed another concept to capture risk, which was the relative risk, through the coefficient of variation which is merely dividing the standard deviation over the mean. So, what you can see here is also the coefficient of variations that each one of these stocks is going to have. Now, as a comment or in spite of the fact that we are going to focus on the standard deviation to capture risk, there are many, many choices here. And this is often called risk measures. And we will provide to you some more reading about that to those of you that are interested in enriching their knowledge about what kind of risk measures you can have when you think about capturing risk in real life situations.

So, just to summarize things, we are going to try and find a good investment plan among these three stocks. And we're going to consider at least two quantities primarily, the expected return, which is the average return that we will calculate as the average of the respective empirical distribution of each one of these stocks. And we're also going to capture risk through the standard deviation, which is going to be equal to the standard deviation of the respective empirical distribution of each one of the stocks. So, let's start first with a simple scenario, in which we are interested in investing our money in exactly one of these stocks. And really what we are worried about is the trade-off between the expected return, that's kind of the upside, and the risk that we would like hopefully to minimize, which is again captured by the standard deviation of the respective distributions. Now, when you look on the data that we have so far, it's very clear that if what you mostly worried about is the expected return, then Amazon looks like the most attractive investment, because it has the highest expected return among all of these three stocks. On the other hand, if what you worried about is mostly risk, you want to minimize your risk, then perhaps you would like to consider Home Depot that has the smallest standard deviation of 5.26%. Now clearly, this is not a 0-1 decision. In this case, it is, but you can have diverse trade-off between the expected risk and the expected return. And we're going to talk more about later. But one thing is clear, I hope that if you only want to invest in one stock, you really have no motivation to consider Walmart as an investment, why? Well, if you look carefully, Walmart is dominated by Amazon with respect to the expected return and at the same time it is dominated by Home Depot with respect to the risk, with respect to the standard deviation.

So, if you only want to invest in one stock, you're really debating between investing in Amazon or investing in Home Depot, in this case. But now, what we would like to ask ourselves, is there any motivation to go beyond investing in one stock? And, what could be the benefits of investing across multiple stocks among the possibilities that we have here: Amazon, Home Depot and Walmart? So, the next question is, what could be possibly benefit from investing in multiple stocks and not in one only? And to understand what could be the benefits of this, we will need to introduce some new concepts and that's what we're going to do next.

Video 3.3: Portfolio Diversification of Stocks: Part One (14:54)

So, what we're going to do next is to introduce two concepts, covariance, and linear correlation, that will help us understand and describe the co-behavior of different random quantities and specifically, different stocks that we have in this particular use case. So, what you see here is again, a plot that describes the co-behavior of the Amazon return and the Walmart return. So, what you see here is a plot that describes the 60 months of historical data that we have. Each point is the returns of Amazon and Walmart in one particular month, where the x value is the return of Amazon, and the y value is the return of Walmart in the same month. So, we have all of these points plotted here. And what we are interested is to measure the numerical and the directional and the degree by which these two quantities covary together. So, before we go and discuss the covariance concept, I would like to start with a reminder about a concept we already discussed, which is the variance, that is applied to one random quantity or to one random variable.

And I'm going to specifically describe this concept with respect to an empirical distribution that has a weight of one over n across and observations, but there is a generalization of that that you can read about in the supporting material. So, what do we do, and how do we calculate the variance in this particular setting? We take each one of the possible observations, measure its square difference from the mean of all the observations. The mean will be x bar, and the square difference will be (x_1 - x bar)^2, (x_2 - x bar)^2, and so forth, and what we do is essentially averaging the square differences across all of these observations, again the square difference of each observation from its mean and that gives us the variance. And what the coherence is, in some sense a generalization of that, right? What are we going to do here? Now, each observation consists of two values, in this case, the Amazon return and the Walmart return, but more generally, if we consider a random variable x and a random variable y, it's going to be x_1, y_1, x_2, y_2 and so forth, right? And what we're going to do now is, for each observation, we're going to take the difference of the x observation from the mean across all the x observations, that's going to be x_1 - x bar for the first observation, and we multiply it by the difference of the y observation from the mean of the y observations, specifically, y_1 - y bar in the first observation and then x_2 - x bar, times y_2 - y bar. And again, what we're going to do is to average all of these products across all the observations, and that's going to give us what we call or denote by s_x, y, which is the covariance of x and y.

Now, just as a sanity check to see why the covariance is a generalization of the variance, if particularly, you're going to take the special case when you covary x with respect to x, you're going to get back exactly the variance definition, right? So, covariance is a generalization of the variance when we apply it to two different random variables, x and y. Now, one thing that is different between the covariance and the variance is that if you think about that, the variance consists of only positive terms, right? Because you always take the square of the difference, and that's always positive. That's not true anymore when you consider the covariance, right? Indeed, some of these products will be positive; specifically, if you take the upper right quadrant where both the y value and the x value are higher than the respective means, that's going to give us a positive contribution towards the covariances. Similarly, if we take the lower left quadrant, right? Where both values are below the mean, right? So, we have two negative values multiplied by each other, that still gives you a positive value. But when you take the upper left quadrant, right? What you get there is that the y observation is going to be above the mean, and the x observation is going to be below the mean, and the product of these two is going to be negative, and similarly, the situation will be similarly, but just flipped when we take the lower right quadrant, right? So, if we think about all of these observations as the points that we see in the plot and we see the x-axis and the y-axis at zero, at the origin representing the mean of the x values and of the y values, what we get is something that consists of both positive terms and negative terms.

And indeed, the covariance is going to be positive if the weight of all the positive quantities is going to out weight the weight of all the negative quantities and vice versa. So, the covariance indeed can get values that are both positive or negative depending on how these points are going to be distributed across the four quadrants, as we saw. And the covariance essentially can be normalized to give us what we call the correlation factor. How do we normalize that? How do we normalize the covariance? We essentially take the covariance and divide it by the standard deviation of the x values and the standard deviation of the y values that give us a unitless quantity that is called the correlation factor that we denote as r_x, y. And what is nice about the correlation factor, that it can get values exactly between -1 and +1. And when the correlation factor is equal to +1, we call this situation that these two random variables are perfectly positively correlated. And similarly, when the correlation factor is -1, we would say that the two random variables are perfectly negatively correlated.

But in most cases, we're going to have different values for any two random quantities that's going to be between -1 and 1, and that's going to be something that we will try to use next. But before that, let's just test our intuition, and I would like you to look on the different plots that we see here, figure A, figure B, figure C, and so forth, that describe co-behavior of two random variables, and what we would like to see is if you have a good intuition to guess what the correlation factor is likely to be, right? Let's just look on the first example, where all the points are aligned on one line, right? This is a situation where exactly where we have perfect correlation. In fact, only in that case, where all the points are lined up on one line, is when we are going to have a perfect correlation. So, being perfectly correlated corresponds to all the points aligned on one line. And if the line has a positive trend, then it's going to be a correlation of +1; if it has a negative trend, it's going to be a correlation of -1. So, this is figure A. And now, take a moment to think about all the other figures that we see here.

So, let's just look at figure B. What you see here is again, points that are aligned on exactly a perfect line, and indeed you have a correlation factor equal to one. And what I would like you to note that this is true both for figure A and for figure B. So, what the correlation factor is not being impacted by is the slope of the line. It really matters whether the slope is going up or down, but the magnitude of the slope does not impact the value of the correlation factor; it's going to be one in both cases. The next two figures C and D are examples where you don't have perfect correlation, but what you can see is that in figure C, there is an upward trend, and the points are almost around one line. So, the correlation will be highly positively correlated close to one, but not perfectly on one line. So, it's not going to be one; it's going to be 0.8. And similarly, as you can see from figure D, you can eyeball that and see that they are about around a negative trend line, and therefore, the correlation is negative, but again, it's not perfectly aligned on one line. And because of that, the value is not -1; it's -0.4.

The last figure might be very puzzling for you, right? Because when you look at it, there seems to be a clear correlation between the points. They don't look like randomly positioned within the space, right? But nevertheless, the correlation factor is equal to zero. And this is a very important thing to remember. The correlation concept, the correlation factor that we discussed, does not capture all correlation forms or all forms of co-behavior between random variables. It really measures only linear correlation. And the linear correlation really depends, again; if you go back to the definition of the covariance and the correlation factor, it really depends on the relative weight of all the points in the different quadrants, right? And in this particular case, there is a complete symmetry around the region that will exactly give us a situation; when you think about the formula of the covariance is going to exactly have a situation when each point will have a point that cancels its contribution and therefore, the total contribution and the covariance is going to be zero and therefore, the correlation factor will be zero. That does not mean that these two random variables do not co-behave in a certain pattern. It just means that that pattern is non-linear, and therefore, might not be able to captured by the concept of linear correlation.

So, again, just to summarize, correlation factor, covariance captured linear co-behavior or to the extent to which two quantities behave linearly together, it does not capture necessarily other forms of co-behavior. And again, you see here more examples that you can test yourself and develop your intuition and then test it against the answers to see whether you got the concept right or not. I don't want to connect that to previous modules, right? And I want to remind you this picture. This is the picture that we saw when we discussed the SIR model, right? And this was part of the output of the SIR model that plotted for us the time until the peak arrives of the infections. This is a graph that described to us the time until the peak infection arrives against the magnitude of the peak, right? And if you look carefully here, we didn't calculate this precisely here, but the intuition that you see here is that you have a negative correlation, not perfectly negatively correlated but somewhat negative, right? And that should maybe match the intuition, right? That if the peak arrives very quickly, then it should be higher, and if it takes longer time for the peak to arrive, then the peak is going to be lower, right? So, to some extent, that might connect to things that you heard with respect to pandemics which is flatten the curve, right? The flatten the curve means that we want the peak to arrive after a longer period of time with the hope that then the peak will be lower, right? As opposed to a situation when the peak arrives faster, and the height of the peak is higher, right?

So, this is exactly capturing that intuition, right? Let's contrast this picture with the picture of the values of the beta and gamma parameters, that again, we model these two independent distributions, right? Uniform and triangular, right? But as you can see here, it's really all over the place, and it's not hard to see that the correlation is really zero here. And indeed, independent random variables have correlation zero, but, by the way, that does not mean, as we saw before from the previous example, that correlation equals zero between two random variables means that they are independent. Again, look on all of these nonlinear co-behaviors that were clearly not independent of each other but still gave us a correlation factor equal to zero, right? So, just to summarize, the two concepts of covariance and correlation factors capture and measure how close to a straight line, how closely linearly correlated data points are, or in more generally to random variables are, right? It does not capture non-linear patterns, as you can see here, where the correlation can still be zero correlation factor, but it just tells us that the linear correlation is zero. It does not mean that these two quantities do not co-behave in a specific form. So, what we want to ask ourselves next is how can we use the concept of covariance and correlation factors to inform the decision about investing in a portfolio of stocks, and we're going to talk about this next.

Video 3.4: Portfolio Diversification of Stocks: Part Two (17:51)

So next, we would like to see how to use the two concepts of covariance and the correlation factors, to inform the decision about how to split or diversify your stock investment across the different stocks that we have been discussing. So, what you see here is the pairwise correlation factors, but in terms of graphs that you can see here, the plots that describe to you how each pair cobe have and their respective correlation factors, calculated just as the same way that we just discussed. And in the table, you can see the different values and it should not be surprising to you that across the main diagonal the value of the correlation factor is one, because indeed when you think about the correlation between random variable to itself, it's always perfectly correlated and equal to one. So, this is again the summary of these values, and the question now is, how to use that to go beyond an investment in a single stock and optimize an investment in a portfolio of stocks? And we will start with a relatively simpler situation, when we're considering the investment in two-stock. And let's see how you can model that in a more formal way.

So, the investment problem that we are thinking about, would be reducing to finding the fraction "a" of money invested in Amazon, whereas the remainder of "1- a" be invested in Walmart. So, we're going to consider Amazon and Walmart now only, and we want to have a decision variable, a decision to invest a fraction a in Amazon, and a fraction 1 - a in Walmart. And again, I want to remind you, these are the means, or expected returns of each one of these stocks, the standard deviation of each one of the stocks and the correlation factor that is equal to 0.11. And just again to formalize what we are doing here, we are forming a model that takes as an input, three quantities. The random variable X that corresponds to the random return of Amazon.

We're going to model that based on the empirical distribution that we created for the Amazon return. Similarly, we're going to have the empirical distribution, y, for the return of Walmart. And we're going to have the decision a, to allocate a fraction of the money in the Amazon stock. And what we're going to get from the model in return is a random variable z, that capture the random return that we're going to get from this portfolio. So z, is it going to be equal to aX + (1-a)Y. And the next immediate question is, to characterize this random variable z, which will be the output of the model. Again, uncertainty in, uncertainty out, as we discussed in the past. What we would like to do is to understand how this random variable z behaves, and specifically what we are interested in is to calculate the expected return of the portfolio that is captured by the random variable z, and the standard deviation, the risk of that portfolio. And well, let's start with something simple, right?

Like if we invest all the money in Amazon, namely, a is going to be equal to one. That's something we can just take from the table that we already calculated because that's going to correspond exactly to the mean of Amazon and the standard deviation of the Amazon return. And similarly, if we decide to consider the value a=0, which means that we invest all our money in Walmart, that's again easy and we're going to be able to calculate that from what we already did. So, let's just now take the middle point a = 0.5. So, we split our money half between these two stocks. And again, what we want to do is to see how we calculate the expected return in this case, and the standard deviation of the return in this case. And just to simplify things at the beginning, I'm going to assume for now that the correlation factor is equal to zero. So, we're first going to discuss that case and then extend it to the situation when the correlation factor is not equal zero, in this case, we know it's equal to 0.11. So, this is somewhat similar to something that we've already discussed. And again, I'm going to assume for now that X and Y are independent random variables. And what I want to remind you, that we already looked on the situation when we looked on a random variable Z, that was the sum of X and Y. That's something we already discussed, and what we saw there is that the mean was the sum of the means, and the variance of Z was the sum of the variances. But that's not what we're trying to do here.

That's maybe a special case of what we are trying to do here. In our case, what we are interested now, Z = aX + (1-a) Y, this is a specific case where we only look on X and Y. So, the linear coefficients between X and Y are 1. Now, we want to consider a more general situation where we look on Z = aX + bY, which is even more general than what we are trying to do, which is a and 1 - a. We're going to generalize that to two coefficients, one before X and one before Y, Z = aX + bY, how do you calculate the mean of Z and the variance of Z? And the mean of Z is going to be very very straightforward, because it's going to be just the linear combination of the respective means of X and Y. Specifically, it's going to be equal to a times the expectation of X plus B times the expectation of Y. And again, to know or to see why, I will be able to read about in the supporting material of the course. When it comes to the variance of Z, we need to be a little bit more careful, and hopefully it's not counterintuitive. Hopefully it's very intuitive to you, why we will need to essentially, square the coefficients a and b. So, specifically look on the variance of Z, VAR[Z] equal to a squared times the variance of X plus B squared times the variance of Y.

And the intuition should be that if you take X and just scale it, the variance of aX is a square time the variance of X, because the variance is always about squaring things. So, again, I want to highlight the fact that a and b here can be any numbers, and specifically you can take a to be equal to 1 to get back to something that we already discussed, but that also captures the situation that we are now considering, when we look on two coefficient a and 1 - a. And let's see now, how to use this insight. And going back to what we are trying to do, which is calculating the mean return, the expected return and the risk of the portfolio, that is resulting by making a decision to invest half of our money in Amazon and half of our money in Walmart. So, the mean again, will be weighting the respective means of Amazon and Walmart by half and summing up, specifically, (0.5) (3.23%) + (0.5) (1.8%). And that gives us a return of 2.15%. Before I tell you the answer with respect to the standard deviation, I would like you to test your intuition.

So, what do we see in this graph? We have, basically, the X axis is going to be the amount of investment that we're going to put on Walmart, that's going to be the value of A. And the Y axis is going to be the resulting standard deviation. And we've already seen that for a = 0, we get the standard deviation of Walmart, and for a = 1, we get the standard deviation of Amazon. And now we are considering the point a = 0.5, and you see here multiple points, A, B, C and D. And what I would like you to do is to test your intuition, what is going to be among these points, the value of investing half of our money in Amazon, half of our money in Walmart. Specifically, what's going to be the standard deviation of that portfolio among these four points? And before we go back to this graph, let's just calculate that. Again, it's going to be the square of half times the variance of Amazon plus the square of half times the variance of Walmart. And that's going to give us a standard deviation of the portfolio, 4.99%. And again, just to remind you, we are still assuming that the correlation is zero. So, let's just go back to where that is positioned, and this is indeed the point that is actually below the line of what we discussed.

So, let's just go back to this graph and you can see here that the value 4.99% is the point C, which is below the point B that was exactly on the straight line, connecting the points of Amazon and Walmart, that you see here. So, what are the insights from this? The insights are that in a portfolio of 50-50, you actually get that the resulting standard deviation of the portfolio is less than the average of the risks of each one of the single stock portfolios, which is the average between point 8.4 and 5.38, right, but it's even stronger than that. The resulting value is not only lower than the average risk of the two portfolios, it's in fact lower than the minimum risk, which is equal to the risk of Walmart, which is equal to 5.38%. So, something very powerful is happening here. We are getting value that reduces the risk from being able to invest in a portfolio of stocks. And if you want to connect that to a similar intuition of something that we've seen in the past we already discussed, this is a go-back to the retail example and the risk pulling effect that we talked about, right? That when you sum together there, there was the sum of random variables that basically reduced the relative risk. Here, the diversification between different random variables is giving us some major benefits with respect to the resulting risk, which is in fact allowing us to reduce the risk beyond the minimum of the two things that we ever joined.

So, this is again a situation that provides major insight because it tells you again that the standard deviation that risk does not behave in a linear manner like the expectation or the mean. And again, another thing that I would like to highlight here that we get this benefit in spite of the fact that the correlation is equal to zero. That's kind of striking because to some extent, we are thinking about the situation when these two random quantities are independent of each other, and you still get that benefit. And what you see in this graph is essentially calculating the same value of the standard deviation for every possible value between zero and one for A, and again, that's under the assumption that the correlation between the Walmart return and the Amazon return is equal to zero. Now, what we know, we know that the correlation factor between Amazon and Walmart and we already calculated that is not equal to zero, it is equal to 0.11. And the next question is, can we actually go beyond the scenario when the correlation is equal zero and do a similar calculation? Well, the answer is yes. But it will require us to modify the formula that we have for the variance of Z.

So, again, this is the situation when Z is equal to a linear combination of X and Y; aX +bY. The mean of Z is just the same as before. The mean is not being impacted by the correlation between the two random variables. So, it's exactly a E[X] + b E[Y], but the variance of Z is modified. It is impacted by the fact that they are now positively correlated or that they are correlated, that the correlation is not equal to zero. So, we have for the VAR[Z] = (a)^2 VAR[X] + (b)^2 VAR[Y] as before, but now we add another term, and again, I'm giving you the term now and you can read why that is the case in the supporting material to the course, but we have to add now a term that is equal to exactly 2 a b SD[X] SD[Y] CORR[X,Y]. And again, these are all quantities that we've already calculated. So, we are well-suited to go ahead and calculate the variance of Z even when the correlation factor is not equal to zero. Because we know the standard deviation of X, we know the standard deviation of Y, we know the correlation factor between X and Y, and we know A and B in our case. So, we are all set to go ahead and calculate that. And if you plug all the values into the formula I just presented, you get that the correlation, you get that the risk of the resulting portfolio that when you invest half of your money on Amazon and half of your money on Walmart and the correlation factor is assumed to be 0.11, you get that the resulting standard deviation of the portfolio Z is 5.23%. So, going back to the picture we've seen before, what you see here is that the standard deviation of the portfolio is now higher than what we had before when the correlation was zero. But nevertheless, it's still lower than the minimum standard deviation between Amazon and Walmart. And if you go back to the initial table that we provided you about the tree stocks, it's also lower than the standard deviation that we saw for Home depot. And this is very interesting insight. Remember, when we discussed investing in only a single stock, we basically said that we should not consider Walmart at all because it was dominated with respect to the expected return by Amazon and with respect to the standard deviation, it was dominated by Home Depot.

But when we're thinking about a portfolio that invests across different stocks, then suddenly in Walmart has its own value because it's co-behavior with Amazon and in that case, there is an interest or there is benefit from investing in Walmart as part of an overall portfolio. And this is kind of a very important insight that you should remember that looking on the marginal distributions of random variables is missing often a lot of important insights, and you have to understand the co-behavior because that's going to give you a lot of benefits and a lot of insights that you don't get when you only consider each one of the random variables separately. So, just to summarize, this is what we call a diversification concept. For two-stock portfolio, this is simple, but you can extend that to a multiple-stock portfolio. Basically, we want to ask ourselves, how does the risk of the portfolio change as the correlation factor between the two stocks decreases. We already saw that it gives us major benefits even when the correlation is zero, and in fact even when the correlation is positive. So, we know already that there is value, but we want now to understand how this value or how the magnitude of the value changes as a function of the correlation factors between the two random quantities. That's what we're going to discuss next.

Video 3.5: Covariance and Correlation (8:54)

So again, next we would like to explore how the variance of a portfolio z, and how its risk specifically, changes as a function of the correlation factor between x and y. Again, assuming that the portfolio consists of investing into stocks in our case. So, let's just go back to how we express the variance of z. So, the variance is a^2 VAR[x] + (1 - a)^2 VAR[y]. And then we have this expression that if we specialize it to our situation is going to be, 2a(1 -

a)SD[x]SD[y]CORR[x,y]. And now observe that all the terms here are positive other than the correlation factor. Which tells us that as the correlation factor decreases the variance of z and therefore the standard deviation of z is going to also be decreasing.

So, less correlation is good for us. And when I say less correlation, that means that zero is better than positive correlation and that negative correlation is better than zero correlation. So, the lower the correlation is the more benefit we get in terms of reduction of the overall risk of the portfolio. Again, as long as we measure it through the standard deviation. So, this is being captured by this diagram, where again, we're going back to calculating the standard deviation of the portfolio for any value of a, that's the X axis. And essentially what you see here are the graphs where the correlation has different values. And this is again reinforcing some of the insights that we've seen before.

Particularly, it tells you that you get benefit even when there is no correlation or even when you have positive correlation. In fact, the only situation when you don't get any benefit of risk reduction is when the two stocks are perfectly positively linearly correlated, namely, when the correlation factor is equal to one. In all other scenarios, you get benefit. Of course, the biggest benefit you get when the two stocks are perfectly negatively correlated, and the correlation factor is -1. And negative correlation factor is better than zero correlation factor and then positive correlation factors. So, that's going to give you some insights about directionally, correlation between stocks is very important to understand. And specifically, when you find correlation that is negative that's very, very powerful for you to reducing risk. But again, you don't need to have negative correlation to reduce risk. Through diversification, you can even have no correlation at all or even positive correlation and you're still going to get some benefits. Okay, so hopefully that's a powerful insight that you can take with you going forward. But let's just go back and think about our portfolio optimization problem, that we posed at the beginning of the discussion. So, so far we were looking just on the risk of the portfolio. And what I would like to do now is to expand that view and look in both the expected return and the risk, simultaneously.

And basically, the graph that you see here is all the resulting portfolios with their respective expected return, the Y value, and the standard deviation. Namely, the risk as the X value. You see all of the returns as a function of a going from 0.0 to 1.0. So again, at 0.0, you get exactly the performance of the Walmart stock alone, and at 1.0, you get the performance of the Amazon stock alone. And what you see all the points on the curve, capturing different values of a between 0.0 and 1.0. And if you're actually interested in minimizing the risk of the portfolio, this is the point that is left most. This is the point that you get when you set up a to be 0.27. That gives you the minimum risk level that you can get by diversifying across these two stocks, and that's equal to 4.75%. But what is important to understand, that there is no way to decide which one of these points is better. They all represent different tradeoffs between expected return and risk. And as you go from a = 0.0 to a = 1.0, you always increase the return. That's increasing essentially linearly between Walmart and Amazon. Amazon has a higher return.

The more you invest at Amazon, you're going to get more return. But pay attention to the fact that the risk does not behave in a similar fashion. The risk is first going down and then starts to go up again. Namely, it's going down, so the curve goes to the left. And then it starts to go back to the right. And that provides you the insight that the diversification is going to allow you to reduce risk. That said, as a manager you have to decide what point you want to actually have on the curve here. And for that, there is no formula for that. That's a managerial choice that you have to make based on your risk appetite as we say. And again, this is sending again the message that analytics can inform and support decisions, but it doesn't waive managers and decision makers of making choices about what matters to them more, risk versus expected return in this case. And based on the balance that they want to have, they can now see all the possible points that will give them the optimal solution with respect to their choice.

Okay, so before we end the discussion about correlation and covariance, I would like to provide you another way to visualize correlation. What you can see on the sequence of graphs here are two sinusoidal graphs, blue and red, and then a third one, which is the sum of the height that is presented in purple, in which case of correlation that is equal to -1.0. Namely, perfectly negatively correlated. When they are completely opposite to each other and essentially, we have no variance. The variance is essentially equal to zero. They cancel each other. Now as you start moving to the right, you start to have a correlation that is increasing from -1.0, let's say to -0.7, and then to 0.0 and then to 0.5 and then to 0.7 and finally to 1.0, when the two curves are perfectly positively correlated. Namely, they are perfectly aligned with each other. And that gives you the highest variability. And again, you can see that as you increase the correlation, the variance, the joint variance, the variance of the sum is going up. And that again provides the intuition that correlation is really an important concept to think and model and capture co-behavior of random quantities. But as I said, that's not the only way or not the only concept to capture co-behavior of random quantities. In fact, we saw already examples in which correlation did not capture some very visible co-behaviors of random quantities. So, it has its own limitations. And what I would like to do next is to talk about a few other concepts that are important in understanding the co-behavior of random variables.

Video 3.6: Module Wrap-Up (1:49)

I would like to summarize the main takeaways from this module. First, I think you can appreciate that risk or uncertainty isn't all bad. It can actually be harnessed to your advantage. And we saw that through the portfolio diversification example where it can actually help you make money, right?

We saw also that correlation is often key in reducing and managing risk and it needs to be understood and considered and leveraged, so don't ignore it, try to understand correlation. And in that context, understand that zero or negative correlation are not necessary to have benefits from diversification. We saw that in fact, unless the correlation is exactly one, you can still get major benefits from correlation. And related to that, risk pulling and portfolio diversification that you can think about them as the linear sum of random variables are powerful strategies to mitigate risk. And finally, the co-behavior of random variables can help you. It can help you by using information obtained on the value of one random variable to update and better understand the conditional probability of another random variable. This is a key concept that enables predictive models and you're going to talk about it more in the coming modules.