EMERITUS
INSTITUTE OF MANAGEMENT

Week 9

Video Transcripts

Video 1 (10:41): Normal distribution

Hello! everyone, in this segment, we'll see examples of continuous valued random variables where the random variable can take any value in the real, in particular, we'll do this through a very important distribution, namely the normal distribution. We'll see many examples and applications of this...this distribution, in this segment.

Let me begin with an introduction to the normal distribution. This is also referred to as the Gaussian distribution, named after Gauss, as many of you may be familiar with. This is one of the most important distribution in the theory of probability and statistics. It arises in many applications, and it has very nice properties. For instance, it provides a very good approximation for sum of large number of uncertain quantities, and we will see a formal statement of this fact later in the course.

Now, this distribution is specified by two parameters: the mean and the standard deviation. So, it's just a two- parameter distribution, and the figure here represents the probability density function of...of the normal distribution. Note that a normally distributed random variable is a continuous random variable that can take any value from negative infinity to positive infinity. So, we cannot really specify a probability of any particular value, like we were able to do in a discrete value distribution. In the case of continuous distributions, the distributions are specified by a density function, like we have in this figure. Let me give a formal interpretation of the density function.

So, suppose the density function for this normal distribution is given by 'f', and 'x' is a random variable that is distributed according to this distribution, then, we can interpret the density function as follows. If we think about the probability of 'X' belonging in a small interval around little 'x' and...and little 'x' plus 'dx', that can be approximated by 'fx' times 'dx'. So, in this figure, suppose this is 'x', and this is 'x' plus 'dx'. The probability that the random variable takes value in this small interval is just the area under this...under this curve, okay! and for 'dx' small, area is approximately 'fx' times 'dx', okay!

So, that's the interpretation of the density function. So, this approximation is good for 'dx' small. In general, the probability that 'X' lies in some interval 'a' to 'b', can be computed as...as the area under the curve between 'a' to 'b', which is basically an integral from 'a' to 'b' of 'fx' '(x)' 'dx'. Now, we know that, probability that 'X' belongs to negative infinity to infinity is '1'. Therefore, the integral from negative infinity to infinity of the density function must be equal to '1'. Another important property I would like to note is, that the density function is symmetric around the mean zero. Therefore, integral from negative infinity to '0' must be half, okay! So, this area is half. As I mentioned earlier, the normal distribution is specified by two parameters, the mean, and the variance, and 'N' mu comma sigma square denotes a...a normal distribution, with mean mu, and standard deviation sigma. A normal distribution with mean zero and standard deviation one, denoted as 'N(0,1)', this is referred to as the standard normal.

Now, let us consider a random variable 'Z', which is distributed according to a standard normal distribution. And, I want to compute probability, that 'Z' is less than equal to '1.3'. So, in this figure, suppose '1.3' is somewhere here, I want to compute the probability that, 'Z' lies from negative infinity to '1.3'. As we just discussed, this is just the area under the curve, from negative infinity to '1.3', so, it is this...this area. Now, this can be computed easily using some-using an Excel or an 'R' function. Now, we don't really need to integrate the density function from negative infinity to '1.3'. There are inbuilt functions in Excel, 'R', and other packages, which can directly give you this probability. Let me illustrate, how to read it off a standard normal table. So, here we have a standard normal cumulative probability table. We have the values of 'Z', and for every value of 'Z' in this table, we have the cumulative probability until 'Z'.

So, we wanted to find the probability that, 'Z' is less than '1.3'. So, we just go down to '1.3', 0, 0, and this is the number. So, the area under the curve until '1.3', is just '0.9032'. So, we can just use this table to compute this probability. This answer we can write as '0.9032'. Now, let's look at the second question. So, I want to find 'Z', such that probability 'Z' is less than equal to little 'z' is '0.95'. So, we can go back to that table again. And now, we want to find value of 'Z', such that the corresponding area under the curve is '0.95'. So, we'll try to find '0.95' in this table, and we have '0.9495'and '9505'. So, probability 'Z', less than equal to '1.64' is '0.9495', and probability 'Z', less than equal to '1.65' is '0.9505', okay! So, the number that we are interested in is somewhere between '1.64' and '1.65'.

So 'Z' is between '1.64' and '1.65'. If we use an Excel, or an 'R' function, we can get a better approximation to the value of 'Z'. While we will never directly use the density function, let me still give you the functional form of the density function, for standard normal. So 'f$_z$' of '(Z)' is one over square root two pi e-power, minus 'x' squared by '2'. And for a general normal distribution, let's say, 'X' is distributed normally with a mean mu, and a variance sigma square, the density function is '1' over square root '2' pi, e-power minus 'x', minus mu, all square by '2' sigma square. Note that, we'll never directly use this functional form of the density function. But, I still wanted to mention this functional form, and also point out that the functional form contains these irrational numbers pi and e. So, computing exact probabilities may not be possible. For instance, if in this question, we may not be able to find the exact 'Z', such that the cumulative probability up to that 'Z' is exactly '0.95', because 'Z' itself may be an irrational number.

Now, of course, we can compute an approximation to 'Z', to any decimal place that we want. Another important fact, that I want to discuss is, that we can convert any general normally distributed random variable, to a standard normal random variable. So, suppose we have a random variable 'X', which is distributed according to a normal distribution, with mean mu and standard deviation sigma. Then, if we consider the following transformation, 'X' minus mu divided by sigma, we get a standard normal random variable, so let's verify this. So, expected value of 'X' minus mu divided by sigma is expected value of 'X' divided by sigma minus mu by sigma, which is '0'. Also, variance of 'X' minus mu by sigma is one by sigma square, variance of 'X', which is '1'.

So, with this transformation of 'X', namely 'X' minus mu divided by sigma, we get a new normally distributed random variable with mean zero, and standard deviation one, which is a standard normal random variable.

Video 2 (4:44): Example 1

Let's see an example of the normal distribution. So, here we have two stocks, 'A' and 'B', with normally distributed returns. Stock 'A' has a normally distributed return with mean '15%' and standard deviation '10%'. So, stock 'B' has a higher return, but also a higher standard deviation, which means it's more risky. Now, I want to compare two portfolios. One is a safe portfolio with '70%' invested in 'A', and '30%' invested in 'B', and a risky portfolio, which has '30%' in 'A', and '70%' in 'B'. Now, I'll use 'S' to denote the random return of the safe portfolio, and 'R' to denote the random return of the risky portfolio. So, let's compute expected value of 'S' and expected value of 'R'. By linearity of expectation, expected value of 'S' is going to be '0.7 mu$_x$' plus '0.3 mu$_y$', which is just '0.7' times '0.15' plus '0.3' times '0.25', which is just '0.18'. Similarly, expected value of 'R' is going to be '0.3 mu$_x$' plus '0.7 mu$_y$', which is just '0.3' times '0.15' plus '0.7' times '0.25', which turns out to be '0.2'. As we would expect, the expected return of the risky portfolio 'R', is greater than the expected return of the safe portfolio.

But, what about the risk, or the standard deviation? So, let's compute the standard deviation or the variance of both the portfolios. Now to compute the variance of both the portfolios, we need to compute the variance of sum of two random variables, and as we discussed earlier, we can actually compute the variance of sum of two random variables. So, if we have two random variables, 'X' and 'Y', that are independent, that needs the correlation coefficient, and the covariance are zero, then variance of 'aX' plus 'bY' where 'a' and 'b' are constant can be written as 'a' square variance of 'X', plus 'b' square variance of 'Y'. Now, for general random variables 'X' and 'Y', with a non-zero correlation, we can compute the variance of 'aX' plus 'bY' as follows, okay! 'a' squared times variance of 'X' plus 'b' squared times variance of 'Y', plus two 'ab' times covariance of 'X' and 'Y'. So, for our portfolio example, let us first consider the case where 'X' and 'Y', which are the random returns of stock 'A' and stock 'B', have a correlation of zero.

And remember, the safe portfolio was '0.7X' plus '0.3Y', and the risky portfolio was '0.3X' plus '0.7Y'. Now, we can compute the variance of 'S' as '0.49' sigma 'X' square, plus '0.09' sigma 'Y' square. Which if we plug in the values of sigma 'X' and sigma 'Y', it turns out to be '0.013'. Similarly, variance of 'R' is '0.09' sigma 'X' square plus '0.49' sigma 'Y' square, which turns out to be '0.045', and let me also write the standard deviations of the returns, so, sigma 'S' is just the square root of variance of 'S', which happens to be '0.0114' and sigma 'R' which is the square root of variance of 'R' is about '0.212'. So as expected, the standard deviation and the variance of the risky portfolio return is higher, than the safe portfolio return. So, we have computed the expected value of the returns of the two portfolios, and the variance of the returns of the two portfolios.

Video 3 (5:59): Normally distributed random variables: distribution of sum

Now, suppose we want to compute the probability that, which of the two portfolios has a higher probability of losing money. Now, this is the event, that probability– that the random return 'S' for the safe portfolio is less than equal to zero, and for the risky portfolio probability, 'R' is less than equal to zero. Now, on the previous slide, we have computed the expected value of the random return, and the standard deviation of the random returns, for both the portfolios. But, to compute the probability that the random return is less than equal to zero, we need to know the distribution of 'S' and 'R'. In particular, I need to know the distribution of the sum '0.7X' plus '0.3Y' less than equal to zero.

And for the risky portfolio probability, '0.3X' plus '0.7Y' less than or equal to zero. So, I need to know the distribution of the sum of two random variables. Recall that both 'X' and 'Y' are normally distributed random variables, and we are working under the assumption that 'X' and 'Y' are independent. Now, to find the distributions of 'S' and 'R', we need to find the distributions of sums of independent, normally distributed random variables, and for that, we'll use a very important property of...of the normal distribution, namely the following. Suppose, we have two random variables, 'X' and 'Y', that are normally distributed, and are independent. Then, we have the following two important facts.

First, a linear transformation of a normally distributed random variable is normal. So, 'aX' plus 'b' is normally distributed, when 'a' and 'b' are constants. And secondly, sum of two independent normally distributed random variables, is also normal. So, 'aX' plus 'bY', where 'X' and 'Y' are independent normally distributed random variables, is also normally distributed. So therefore, 'Z' will have a normal distribution, with mean 'a' mu 'X' plus 'b' mu'Y' and variance 'a' square sigma 'X' square 'b' square sigma 'Y' square.

The figure here illustrates, that the density of sum of two independent normal variables is also normal, okay! Now, this is a very important property of the normal distribution, that allows us to compute the distribution, of sum of two or more independent, normally distributed random variables. In general, it is quite challenging to compute the distribution of sums of random variables for general distributions.

Now, with this property of the normal distribution, let's go back to the previous slide, and compute the probability of losing money in both the portfolios. Now, based on what we just discussed, both 'S' and 'R' are normally distributed. Now, recall the expected return of the safe portfolio was '0.18' and variance was '0.013'. Therefore, this distribution is normal '0.18', '0.013'. Now, for the risky portfolio, the expected return was '0.22', and the variance was '0.045'. Therefore, 'R' is distributed normally with mean '022'and variance '0.045'. Now, let's compute the probability that 'S' is less than or equal to zero. Now, we can transform 'S' to a standard normal variable using the transformation that we discussed earlier. So, this probability is going to be same as probability 'S' minus 'mu$_S$' divided by 'sigma$_S$' is less than minus 'mu$_S$' by 'sigma$_S$'. Now, this is a standard normal variable 'Z', and plugging in the...in the values of 'mu$_S$' and 'sigma$_S$', we get this as probability of standard normal variable 'Z' is less than '-1.579'. Similarly, probability that 'R' is less than equal to zero can be written as probability 'R' minus 'mu$_R$' divided by sigma$_R$ is less than minus 'mu$_R$' by 'sigma$_R$'. Now, this again is a standard normal variable 'Z', so this probability can be written as probability 'Z' is less than equal to plugging in the values, we get '-1.038'. Now, '-1.579' is less than '-1.038'. Therefore, probability 'Z', less than equal to '-1.579' is less than probability 'Z' less than equal to '-1.038'. Therefore, probability of losing money is higher in the risky portfolio. Note that, in this example, we were able to compare the probability of losing money in the two portfolios, without even computing the probability of losing money in either portfolio. And this was possible because, we transformed the probability computation into a probability over a standard normal variable for both the portfolios.

So, this transformation, to the standard normal variable is very useful, when working with normal distributions.

Video 4 (4:19): Example 2

Let's consider another example, where we want to compare the following two portfolios; portfolio one, which invest '80%' in stock 'A' and '20%' in stock 'B'; and portfolio two, which invest '90%' in stock 'A' and '10%' in stock 'B'. So, portfolio one is the one that is investing a higher fraction in the risky stock 'B'. We want to find which portfolio has a higher probability of losing money. So, let's first compute the expected value and variance of the returns for both the portfolios. So, expected value of the return of portfolio, 'P_1', I refer to it as 'mu$_1$', is just '0.8 mu$_X$', plus '0.2 mu$_Y$', which if I plug in the numbers, turns out to be '0.17'.
And variance of return of 'P_1' is...is let me denote it by, 'sigma$_1$' square, is '0.64 sigma$_X$' square plus '0.04 sigma$_Y$' square, which if I plug in the number, turns out to be '0.01'. So 'sigma$_1$' is '0.1'. Now, using the property that some of independent normally distributed random variables is also normally distributed, we know that the return of portfolio one is normally distributed, with mean '0.17' and variance '0.01'. So, probability that the return of portfolio one, let me call that 'P1', is less than equal to zero, is just transforming it to a standard normal variable, which is 'P_1' minus 'mu$_1$' divided by 'sigma$_1$', is less than minus 'mu$_1$' by 'sigma$_1$', is just probability that 'Z' is less than equal to '-1.7'. Now, let's do the same exercise for portfolio two.
So, computing the expected value of the return, 'mu$_2$', is just '0.9 mu$_X$' plus '0.1 mu$_Y$', which turns out to be '0.16'. And the variance 'sigma$_2$' square, is just '0.81 sigma$_X$' square plus '0.01 sigma$_Y$' square, which, plugging in the values, we get, '0.009'. So, the distribution of return of portfolio two is just normal '0.16'and variance '0.009'. Now, we can compute the probability that the return of portfolio two is less than equal to zero as probability of the standard normal variable is less than negative 'mu$_2$' by 'sigma$_2$', which if I plug in the values, turns out to be probability 'Z', is less than equal to '-1.68'. So, now '-1.7' is less than '-1.68', which means that probability 'Z' less than equal to '-1.7' is strictly less than probability 'Z' less than equal to '-1.68', which actually means that probability that portfolio one loses money is strictly smaller, than probability portfolio two loses money. This is a bit surprising because portfolio one had higher fraction of the risky stock 'B', and a higher standard deviation of variance '0.01' as compared to variance of portfolio two risk return, which was '0.009'.
So, this example illustrates, that the standard deviation of the return being higher doesn't necessarily imply that probability of losing money is also higher.

Video 5 (5:42): Joint distributions

In the two examples so far, we have assumed that the returns of stocks 'A' and 'B' are independent of each other, and that allowed us to compute the distribution of the returns of the portfolios that we have considered in our examples. Now, what if there was a correlation between the returns of stocks 'A' and 'B'? Can we then compute the distribution of returns of the portfolios that combine stocks 'A' and 'B'? Now, to discuss that, we need to introduce the concept of joint distributions. So, when two random variables are not independent, we need to specify the probability of joint events. So, for instance if 'x' and 'y' are correlated, then we want to specify the probability that 'x' is in some interval 'x', and 'x' plus 'dx', and 'y' is in some other interval, 'y' and 'y' plus 'dy'. And this probability is specified by the joint density function.

The joint density function 'f' that specifies the joint density of two correlated random variables 'x' and 'y' is a function from R_2 to 'R', and the interpretation is the following. Probability that 'X' belongs to this interval little 'x' to little 'x' plus 'dx', and 'Y' belongs to the interval little 'y' to little 'y' plus 'dy' is given by 'f(xy)' times 'dx' times 'dy'. where 'f(xy)' is the density at little 'x' comma little 'y'. As with the probability density function of a single random variable, this density function also satisfies the following two properties, it's non-negative everywhere, and the double integral over 'x' and 'y' must be equal to one. Also, probability of any event, namely that the pair of values 'X' comma 'Y' belong to some set 'B' where 'B' is a subset of R_2 can just be computed by this double integral over 'x' and 'y' over the set in 'B'. Now, given this joint density of 'x' and 'y', we can compute the marginal distribution of 'x' or 'y', and the marginal density can be computed as follows, for instance, marginal density of 'X' can be computed by integrating the joint density over all values of 'Y'. That would be just a function in 'x', and that's just the density function of the random variable 'x'. Similarly, if I was interested in computing the marginal density of 'y', I would just integrate over 'x' 'f(xy)' 'dx'. Now, if 'X' and 'Y' are independent, then specifying just the marginal densities gives me the joint density as well, because the joint density in that case is just the product of the marginal densities.

Now, suppose we have two random variables 'X' and 'Y', that are jointly, normally distributed, and by jointly normally distributed, we mean that the joint density function is as shown in this figure, has this bell shape. Then, the sum of these random variables is also normally distributed. As I mentioned earlier, this is a very important property of the normal distribution, that allows us to compute the distribution of sums of normally distributed random variables, if they're jointly normal. Now, let me also specify the precise functional form of a joint normal density. Suppose mean of 'X' is given by 'mu$_X$', and mean of 'Y' is given by 'mu$_Y$'. Then let me define this vector mu, which is the vector of the two means, 'mu$_X$' and 'mu$_Y$', and let me define this covariance matrix sigma, which is a two by two matrix, where the diagonal elements are just a variance of 'X' and 'Y'. And the off diagonal elements are just the covariance of 'X' and 'Y'. This is referred to as the covariance matrix.

Note that if 'X' and 'Y' were independent, then covariance would be zero, and then the covariance matrix is just a diagonal matrix. Now, given the mean vector and the covariance matrix, we can specify the functional form of the density as follows. So, let's 'z' be the tuple 'x' comma 'y'. Then the density at 'z' can be written as follows. It's one over square root of two pi determinant of sigma, times e-power minus half 'z' minus mu transpose sigma inverse 'z' minus mu. Note that in this example, we specified the density for a Joint distribution of two random variables. Now, if we have 'n' random variables, that are jointly normal, the same functional form follows, only thing changes is the covariance matrix is now 'n' by 'n', and the mean vector is a n-dimensional vector. And exactly the same functional form is the joint normal density for 'n' jointly normally distributed random variables.

Video 6 (6:37): Portfolio with correlated stocks

Now, let's consider the case where returns of stock 'A' and stock 'B' are not independent. In fact, let's consider the case that they are positively correlated and the correlation between the returns is '0.1'. And we want to answer the question, can we construct portfolios that are better than the two portfolios that we considered earlier, the safe portfolio and the risky portfolio? Now, to answer that question, we first need to decide what do we mean by better? Is it higher expected return? Is it lower standard deviation? So we...we need to first define what do we mean by better? Now, the notion of better actually depends on the risk preferences of the investor. A risky investor may be more biased towards higher expected returns, whereas a risk averse investor might be biased towards low standard deviation. So, what we can do is just analyze the expected return, and standard deviation tradeoffs for different proportions of wealth invested in stocks 'A' and 'B'. In particular, we'll consider investing 'w' fraction of wealth in stock 'A', and '1-w' fraction in 'B', for all values of 'w' and zero to one, and then look at the expected returns and standard deviation of all of these portfolios. Let 'P$_w$' be the random return when you invest 'w' fraction in 'A'. And one minus 'w' fraction in 'B'. So that's going to be 'w' times 'X' plus '1-w' times 'Y'.

Now, expected value of 'P$_w$' is just 'w' 'mu$_x$' plus '1-w' 'mu$_y$', and standard deviation or variance of 'P$_w$' is going to be 'w$_2$' 'sigma$_{x2}$' plus '1-w$_2$' 'sigma$_{y2}$'. And since returns of 'x' and 'y' are correlated, it's going to be two times 'w', one minus 'w', covariance of 'x' and 'y', which can be written as 'w$_2$', 'sigma$_{x2}$', plus '1-w$_2$', 'sigma$_{y2}$' , two 'w' '1-w' 'sigma$_x$', 'sigma$_y$' times '0.1'. Let's plot the expected return and standard deviation for the portfolios for different values of 'w'. So, on the 'Y' axis, I have the expected return of the portfolio. And on—and the 'X' axis, I have the standard deviation of portfolio for different values of 'w'. Now, each point on this plot is a portfolio. For instance, the extremes 'w' equals zero and 'w' equal one are investing purely in 'B' or purely in 'A'.

So, if you invest only in 'B' which is 'w' equals zero, the expected return is high; but at the same time, the standard deviation is also high. Now, on the other hand, if you put everything in 'A', that is, 'w' equals one, then the expected return is low, and the standard deviation is also low. Now, suppose you are a risk averse investor, who puts a lot of weight on low standard deviation, that means you want to minimize the standard deviation. You would want to put more money in stock 'A'. Now, is it optimal to put all your money in stock 'A'? Seems natural, because stock 'A' has a lower standard deviation, and is positively correlated with stock 'B'. Surprisingly, this plot suggests otherwise. In fact, we have portfolios here, which invests a non-zero fraction in stock 'B', and have a higher expected return, than just putting everything in stock 'A', and lower standard deviation.

So even though the returns of the two portfolios are positively correlated, diversification helps in having a higher expected return and a lower standard deviation. So, on this slide, we have the expected return and standard deviation of all the portfolios for different values of row, which is the correlation between, returns of stock 'A' and stock 'B', which is just defined as covariance of 'X' and 'Y', divided by 'sigma$_x$' 'sigma$_y$'. Note that the value of diversification decreases, as the correlation increases, but as we showed in the previous slide, even if the returns are positively correlated, there is value in diversification. Of course, the value of diversification is the highest, when the returns of stock 'A' and stock 'B' are negatively correlated. As we can see in this figure, if we think about the correlation of stock 'A' and stock 'B' to be minus one, that is, they are perfectly negatively correlated, the value of diversification is the highest.

This is shown in the blue plot here. So, in the case where the returns of stock 'A' and stock 'B' are perfectly negatively correlated, we can achieve a much higher expected return, at a small standard deviation. As an example, let's compare the case of row being minus one, and row being minus point five. Now suppose, I am risk averse investor, who wants to find a portfolio with the smallest possible standard deviation. In the case, where the correlation between returns of stock 'A' and stock 'B' is one point five, I would choose the following portfolio, okay, so that achieves the following expected return and has this standard deviation. Now, to achieve the same standard deviation for the case where row is minus one, we can get much higher expected return, okay. So, this would be the expected return, for the same standard deviation, if row was minus one. So this illustrates that the value of diversification is higher, if the correlation is more negative. But, it is important to note that diversification helps even when the correlation is positive.

Video 7 (2:04): Markowitz portfolio - efficient frontier

In the examples, we have discussed so far, we have only discussed diversification over two underlying stocks, stock 'A' and stock 'B' in our examples. But in reality, we would like to build a portfolio that diversifies over many different underlying stocks, with that the computation of the best portfolio becomes a bit more challenging, but we can still do it efficiently. In fact, Markowitz did that in his influential 1952 paper where he introduced portfolio theory, and a very important concept of efficient frontier. This paper was very influential and in fact, he got Nobel prize for...for that paper. We'll come back to this concept of efficient frontier, when we are discussing optimization models and formulations later in the class, but let me illustrate this concept through this figure. So, in this figure, again on the 'Y' axis, I have the expected return, and on the 'X' axis is the standard deviation, and the shaded region represents the expected return and standard deviation, for different possible diversifications. Now, the efficient frontier in this example is the following boundary. Now, it's clear, no matter what your risk preferences are, you should never choose a portfolio outside of this efficient frontier.
To illustrate, let's consider a portfolio that is outside this efficient frontier. So suppose, I consider this portfolio in red, now I can achieve the same standard deviation and a better expected return, by choosing this portfolio in green, which is on the efficient frontier. So, portfolios that are outside of this efficient frontier are dominated by portfolios in the efficient frontier, if I'm just focused on expected return and standard deviation.

Video 8 (5:56): Value at risk and summary

Next, I want to introduce another important concept, namely, value-at-risk. This is a measure of risk. So far, we have used standard deviation as a proxy for risk, but we can define other measures of risk for instance, value-at-risk. Let me define this through an example. Suppose, I want to define '99%' value-at-risk for an investment. It's that value of the return, such that probability that your returns will actually be lower than that value is at most '1%'. More specifically, let's consider this example, where we want to compute '99%' value-at-risk, also denoted as, 'VaR' for 'S&P 500'. Now, suppose the annual return of 'S' and 'P' is normally distributed, with mean '8.79%' and standard deviation '15.75%'.
So, if I just want to draw the density, that's going to be normal; this mean is '8.79%'. Now, '99%' value-at-risk is that value 'x', such that the probability that return actually falls below 'x' is only one percent, which means this...this area under the curve must be '0.01'. So, if 'x' is such, that the area under the curve of two 'x' is '0.01', then 'x' is '99%' value-at-risk. Now, as you can see from the definition of value-at-risk, this measure bounds the probability of a loss. So, if I can find and specify you '99%' value-at-risk,

probability that your return will be lower than that value, is bounded by 0.01. On the other hand, standard deviation was more of an aggregate measure of risk. So, in some sense value-at-risk is a more risk averse measure as compared to standard deviation, and this is very commonly used, especially in institutional investments settings such as banks, where you want to find a threshold on the probability of a loss below certain quantity.

Let me illustrate this measure through another simple example. Suppose, you are managing a portfolio that is worth '$100 million', and your average daily payoff is '$0M', with a standard deviation of '$3M', and suppose this payoff is normally distributed, and I want to find '97.5%' value-at-risk for one day. So, let's first plot the distribution of the payoffs– the daily payoffs. We know the mean is zero, and the payoffs are normally distributed. So, we have the nice bell curve as the distribution, and we want to find, '97.5%' one day value-at-risk. So, really we want to find this value 'x', such that the area under this density up to 'x', so all this area is '2.5%'. And so, this area must be '0.025'. So, if 'x' is such that the area under the curve up to 'x' is '0.025', then from the definition of value-at-risk, we know that probability that we will lose 'x', or more is only 0.025, that means, on only '2.5%' of the days, we would lose 'x' or more.

So, this measure of value-at- risk, really defines a threshold with a specified probability. For instance, in this example, we wanted to find '97.5%' value-at-risk, so we found the threshold 'x', such that the probability that payoff is more than 'x' is '97.5%', or in other words, probability that the payoff is less than 'x', which means loss is more than 'x', is '2.5%'. So, this is a more risk-averse measure as compared to standard deviation, which does not have any such probabilistic guarantees. Let me now summarize, what we have seen so far for the normal distribution. This is a very important distribution, and we have seen many examples and applications for this distribution. And we have seen many important properties for this distribution. In particular, we saw that if we have a general normally distributed random variable, it is always useful to convert it into a standard normal, using this transformation. In fact, we saw examples where after this transformation, we did not even need to compute the probabilities, when we were comparing two portfolios.

A very important property that we saw about normally distributed random variables are, if you have two random variables that are jointly normally distributed, then any linear transformation of a normally distributed random variables is also normally distributed, and more importantly, some of jointly normally distributed random variables is normally distributed. Now, this is as I have mentioned before, is a very important property about normally distributed random variables. In general, thinking about the distribution of some of random variables which are correlated, that's a very challenging problem, but if we are just interested, in expected value or variance of sums of the random variables, then we can use these formulas which are... which are consequences of linearity of expectation. So, these are useful formulas that you should keep in mind.

Video 9 (10:18): Bernoulli and binomial distributions

In this segment, we will see many more examples of important distributions, both discrete and continuous. In particular, we will see Bernoulli distribution, binomial distribution, and geometric distribution. They are examples, of important discrete valued distributions. We will also see several other important examples of continuous value distributions, namely the exponential distribution, Poisson distribution, and Poisson process, and lognormal distribution. Let us start with discussing the Bernoulli distribution. This is a very simple discrete value distribution with only two possible outcomes. So anytime we have a random process that has only two possible outcomes, for instance, a success or a failure, we can model it using a...a Bernoulli distribution.

A Bernoulli distribution is a single parameter distribution where we just need to specify the probability of one of the outcomes. Also, any associated random variable will have only two possible values, because there are only two possible outcomes. So for instance, we can consider a random variable 'X', which takes value '1' with probability 'p' in the event of success and '0' with probability '1' minus 'p', denoting the outcome of failure. For that random variable, we can define the cumulative density function as follows, and it's easy to note, that the expected value of this random variable is going to be 'p' and the variance is going to be 'p' times '1' minus 'p'. There are many examples where a Bernoulli distribution would arise naturally. For instance, in display advertising, suppose I want to model whether an ad was clicked or not, okay! So, the probability of clicking an ad, that's the probability of success and this can be modeled using a Bernoulli distribution.

Now, probability that stock price would go up in the next period, this can be modeled using a...a Bernoulli distribution. So, Bernoulli distribution can arise in many natural settings but really Bernoulli distribution is a building block for more richer distributions, as we'll see later. There are many important discrete value distributions, for instance, a binomial distribution, a geometric distribution, or a negative binomial distribution for which a Bernoulli distribution is the basic building block. For instance, a binomial distribution models the number of successes in 'n' independent trials, where each trial can be modelled using a Bernoulli distribution. A geometric distribution models number of trials needed for the first success, where each trial itself is a Bernoulli distribution. A negative binomial distribution models the number of trials before the exit success.

So, for all these distributions, Bernoulli is the basic building block, and we'll discuss this in more detail, in the later slides. Let me begin by describing the binomial distribution. Recall, this was a distribution that models the number of successes in 'n' independent trials, where each trial was actually a Bernoulli event. I'll describe this distribution through an option pricing model. In particular, I consider a binomial evolution of the price of the underlying asset or...or stock. So, let's say, we consider this asset or stock 'S_0' and in one period the price could go up by a factor 'u' or down by a factor 'd'. So, we have these two events, with probability 'p', the price in one period goes to 'u' 'S_0', or with probability '1' minus 'p', the price goes to 'd' 'S_0'. Similarly, after the second period, we have the following three possibilities of the price, so it could be 'u' square, 'ud', or 'd' square of 'S_0'. Note the price 'udS_0' can be achieved in two ways, the price could go up in period one and down in period two or it could go down in period one and up in period two and these possibilities are modeled using the binomial distribution.

Let's consider one of the possibilities, let's say 'u^2dS_0', so the price has gone up in two periods, and down in one period. This can happen in multiple ways, we need to select one period out of three, where the stock price goes down and in the other two, the stock price has gone up. So, this can be done in three choose two ways, where I select the two periods where the price has gone up, which is equal to three. And, the probability of this event is going to be three 'p' square, times '1' minus 'p'. Now the binomial distribution model's, exactly this, probability of each of these four outcomes. More formally, a binomial distribution specifies the distribution over number of successes in 'n' independent trials, where each trial is a Bernoulli event, which has a success probability of 'p' and a failure probability of '1' minus 'p'. Now, an associated random variable, it can take values from zero to 'n'.

So, there are 'n' plus '1' possibilities create values. Probability that the random variable has value exactly 'k', that means there are exactly 'k' success out of 'n' trials is going to be 'n' choose 'k', 'p' raise to the power 'k', '1' minus 'p' raise to the power 'n' minus 'k'. So, we can break this probability down into three components, choosing 'k' of the trials out of 'n' where success happens, so that is, 'n' choose 'k' ways to do that. Probability that success happens in those 'k' chosen trials, that's 'p' is to the power 'k' and the remaining 'n' minus 'k' are failures, which is '1' minus 'p' raise to the power 'n' minus 'k'. Now, using the definition of the expected value and the variance, we can find that the expected value of this random variable is going to be 'n' times 'p' and the variance is going to be 'n' times 'p' times '1' minus 'p'. Recall, we can compute the expected value of 'X' as some over 'k' going from '0' to 'n', probability 'X' equals 'k', which is specified by that binomial formula up there, times 'k', and variance of 'X' is going to be expected value of 'X' square minus expected value of 'X' whole square.

So, we can plug in the probability formula up on the slide, and compute both expectation and variance, but there's also an easy way to compute expectation and variance, in fact, let me think of the random variable 'X', which has a binomial distribution as a sum of Bernoulli distributed random variables. In particular, let me say 'X' is 'Y_1' plus 'Y_2' plus Y_n', where each 'Y_j' is '1' with probability 'p', and '0' with probability '1' minus 'p'. Okay! So each 'Y_j' is a Bernoulli random variable that models success with probability 'p' and failure with probability '1' minus 'p'. In this way, clearly 'X' models the number of successes in 'n' trials. Now, from the discussion about the Bernoulli distribution, we know that expected value of 'Y_j' is equal to 'p' for all 'j', and variance of 'Y_j' is 'p' times '1' minus 'p', and 'Y_j's are independent of each other.

Now, we can use linearity of expectation to compute the expected value of 'X', that's just going to be sum over 'j' going from '1' to 'n', expected value of 'Y_j', and expected value of 'Y_j', is 'p' for all 'j'. That means expected value of 'X' is 'n' times 'p', which is exactly what we had on the previous slide. Similarly, variance of 'X' is going to be sum over 'j' going from '1' to 'n', variance of 'Y_j' and remember, in general, we'll also have covariance terms but since 'Y_j's are iid. The covariance is zero and therefore, the variance of 'X' is 'n' times 'p' times '1' minus 'p'. So, here we see explicitly that Bernoulli is a building block for the binomial distribution. Another very important property that I want to mention about the binomial distribution is, suppose 'n' is sufficiently large, then we can approximate, a binomial distribution by a Gaussian distribution, appropriate mean and appropriate variance. In particular, with a normal distribution with mean, mu is 'np', and variance being 'n' times 'p' times '1' minus 'p'. Now, to understand this approximation, I want to go back to writing a Bernoulli random variable as sum of 'n' Bernoulli random variables. Okay! As we have discussed in the past, sums of independent random variables and in this case, iid.

Bernoulli random variables, they can be very well approximated by a Gaussian distribution. The approximation gets better as 'n' grows large and in fact, for 'n' tending to infinity, the approximation is exact.

Video 10 (2:46): Geometric distribution

The next distribution I want to discuss is the geometric distribution, which essentially models the number of trials until first success, where, again, each trial is a Bernoulli event. So here again, Bernoulli is a basic building block. Now, an associated random variable can take any integer value from zero to infinity. So, 'X' can have value '1', '2', '2', and so on until infinity, and probability that the value of this random variable is exactly 'k', which means there were 'k' trials until first success, is one minus 'p' raised to the power of 'k' minus '1', times 'p'. So, if there are 'k' trials until first success, the first 'k' minus '1' trials must be failures, and therefore, you have '1' minus 'p' raised to the power of 'k' minus '1', and the last trial being a success which gives us the product of 'p' times, '1' minus 'p' raised to the power of 'k' minus '1', and the CDF is given as follows.
Remember, CDF is just the probability that 'X' is less than equal to 'k', which is '1' minus probability 'X' is greater than 'k', which is exactly equal to '1' minus the first 'k' trials are failures. Now, using the definition of expectation, and variance, and plugging in the formula for probability, we can compute the expected value and the variance as follows. So, expected number of trials is going to be '1' over 'p' and variance is going to be '1' minus 'p' divided by 'p' square. In many examples, where 'p' is small, we can approximate this by '1' over 'p' square. So, we see again that a Bernoulli distribution was a building block for the geometric distribution. Let's consider an example of a geometric distribution.
So, let's consider a basketball player who has a '60%' chance of making a free throw, and suppose all free throws are independent. I want to compute the probability that his first successful throw comes on at his third try. So, that's going to be, for the first two throws, is a failure times 'p', where 'p' is '0.6'. So, this is going to be '0.16' times '0.6' is going to be '0.096'.

Video 11 (8:25): Exponential distribution

Next, I want to discuss examples of continuous value distributions. And let me first start with the example of exponential distribution. This is a very important distribution that arises in many modeling scenarios. Let me begin by specifying the PDF and the CDF of this distribution, before discussing its properties. In particular, the PDF for this distribution can be specified as follows. So, this is a positive value distribution and f of 'x' is lambda 'e' to the power minus lambda 'x'.
So, this is a single parameter distribution, so the parameter being lambda, and is nonzero only for nonnegative 'x'. Now, this cumulative density function, or the CDF, which is probability, the random variable 'X' is less than or equal to 'x' is given by '1' minus 'e' power minus lambda 'x', for any 'x' nonnegative. Now, given these PDF and CDFs, we can compute the expectation is '1' over lambda, and variance is '1' over lambda square. Recall, we can compute the expectation of a continuous valued random variable using an integral. So, the expectation of 'X' can be defined as 'x' going from '0' to infinity. 'x' times 'f' of 'x', 'dx', which is '0' to infinity 'x' lambda 'e' power minus lambda 'x', 'dx'. And, using integration by parts, you can see that this is going to be '1' over lambda.
Similarly, variance of 'X' can be computed as expectation of 'X' square, minus expectation of 'X' whole square. So, this is going to be integral, from '0' to infinity, 'x' square, lambda 'e' power minus lambda 'x', 'dx', minus '1' over lambda square. And again, using integration by parts, we would get '1' over lambda

square. So, if the standard deviation for 'X' is going to be '1' over lambda. So, you can note that both expectation and the standard deviation are the same. As I mentioned earlier, this is a very useful distribution, and satisfies many important properties. So, let me give some intuition behind the properties of this distribution. Let me begin by describing an important relation between an exponential distribution and a geometric distribution. In particular, the exponential distribution is a continuous analog of geometric distribution.

Recall, the geometric distribution was a distribution on the number of trials until first success, where each trial was an independent Bernoulli trial with a success probability 'P'. Suppose, 'Y' is a geometric random variable with success probability 'P' in every single trial, then probability 'Y' is bigger than 'k' is '1' minus 'p' raised to the power of 'k', that means the first 'k' trials must be a failure. Now, consider an exponentially distributed random variable 'X', which is 'Exp' of lambda, then probability 'X' is bigger than 'k' is '1' minus 'f' of 'k', which is 'e' power minus lambda 'k'. Okay! And now, let me show you a connection between this formula and this formula. As I said earlier, the exponential distribution is a continuous analog of the geometric distribution. To understand this connection, consider the following setting.

We divide the whole time into intervals of length delta, and think of delta as a very small number. Okay! So you have time, and you divide this whole time into small intervals of length delta each, okay! In each interval, we toss a coin whose success probability is lambda delta and failure probability is '1' minus lambda delta. Okay! If we see a success, we stop, if we observe a failure, we continue. And the value of the random variable is going to be whenever we stop. Now, let me consider random variable 'z' which denotes the stopping time. Now suppose, I'm interested in the probability that 'z' is bigger than 't'. So, 'z' would be bigger than 't' if in all previous intervals of each of length delta, I...I have observed a failure. How many intervals are there until time 't'? 't' by delta.

In all of those intervals, I should have observed a failure, and all of these trials were independent, so this probability is '1' minus lambda delta raise to the power 't' by delta. This probability as limit delta goes to '0', is going to be 'e' power minus lambda 't', which is exactly the same as the exponential distribution probability. So, this explains a very important connection between exponential and the geometric distributions. In fact, we can think about exponential distribution as a geometric distribution, where the Bernoulli trials are happening in continuous time. In every interval of delta width, there's a Bernoulli trial with success probability lambda times delta.

So, this continuous Bernoulli trial interpretation of the exponential distribution is very helpful when trying to think about properties of this exponential distribution. One of the very important properties of the exponential distribution is, that it is memoryless. In particular, suppose I have a random variable 'T', which is exponential with parameter lambda, then probability 'T' is bigger than 's' plus 't', given that 'T' is bigger than 'S'. Same as probability 'T' is bigger than 't'. Let's see two explanations for this, okay! Let's go back to the definition of this conditional probability that would be probability 'T' is bigger than 'S' plus 't'. And 'T' is bigger than 'S' divided by probability 'T' is bigger than 'S', Which if I plug in the CDF—one minus the CDF, I would get 'e' power minus lambda 't', which is exactly the same 'S' probability 'T' bigger than 't'. We can also conclude about this memoryless property from the continuous Bernoulli trial interpretation of the exponential distribution.

So, this is the conditional probability we are interested in. I want to condition on the event that 'T' is bigger than 'S'. That means in the continuous Bernoulli trial interpretation, all the Bernoulli trials until time 'S' have failed. Now condition on this event, I want to find the probability that 'T' is bigger than 'S' plus little 't'. That means that the Bernoulli trials must fail for 't' more units of time. Since, all the Bernoulli trials are independent, this is exactly the probability that 'T' is bigger than little 't', which is precisely the memoryless property. So, we see that this continuous Bernoulli trial interpretation of the exponential distribution helps us derive this very important property of the exponential distribution, namely the memorylessness.

Video 12 (3:06): Example of exponential distribution

Let's see some examples where we use exponential distribution to model. One of the very important applications of exponential distributions is modeling inter-arrival times and an arrival process. As an example, consider a random arrival process of people at a bus station, and suppose the average number of people arriving in an hour is three. In many settings, it's very reasonable to model inter-arrival times as exponential distributions. So, let's do that in this example. So, the inter-arrival time is 'X', which is exponentially distributed with the rate lambda, since the average number of people in an hour is three, expected value of 'X' must be '1' by 3', which is the expected length of the inter-arrival time, which is '1' over lambda.

So, lambda is '3', and I am interested in the probability that the next person arrives in less than one hour. So, the inter-arrival time is less than one hour, which is probability 'X' less than equal to '1', which is going to be '1' minus 'e' power minus lambda, which is going to be '1' minus '1' over 'e' cube. Let's discuss another very important application that of a call center. Now, in a call center, we need to model several different random processes. First, the arrival process. How the calls arrive to the call center? What is their rate? Second, the service process. How the calls are serviced, or how much time it needs to service one call? And third, the abandonment process.

Not all calls are serviced as soon as they arrive to the call center. The calls, which are not serviced immediately, queue, and wait for service. Now, some of these, due to impatience or other factors, may abandon the system, and we need to model the time to abandonment. Now, in this application, exponential distribution is very commonly used, in the arrival process, the inter-arrival time is modeled using an exponential distribution. For the service process, the service time is model using an exponential distribution, and also for the abandonment process, the time to abandonment is also modeled using a exponential distribution.

So, you can see that exponential distribution shows up in many important applications. We will see later that modelling these random processes using exponential distributions makes the model tractable. Here, I also want to introduce something called a Poisson process. So, any arrival process where inter-arrival times are exponentially distributed is a Poisson process.

Video 13 (5:24): Poisson distribution

Let me introduce the Poisson distribution. Now, Poisson distribution is a discrete value distribution that models the number of events occurring in a fixed time interval. A random variable distributed according to a Poisson distribution can take values from zero to infinity, any integer from zero to infinity. And, probability that the value is exactly 'k' is given by the following formula, 'e' power minus lambda, lambda to the 'k' divided by 'k' factorial. So, the Poisson distribution is also a single parameter distribution where lambda is the parameter. And, lambda is essentially the average number of events in that fixed time interval. So, expected value of 'X' is going to be lambda and variance is going to be lambda.

Let's discuss a small multiple choice question. Which of the following is most likely to be well modelled by a Poisson distribution? So, the first choice is number of trains arriving at a station every hour. This is the precisely the setting very well modeled using a Poisson distribution. Let's look at the second choice. Number of lottery winners each year that live in Manhattan. Now, if you think about this option for a bit, it is more likely to be well modeled using a binomial distribution. Suppose, each person in— living in Manhattan has some probability of winning the lottery. Therefore, the number of lottery winners would be modelled by using a binomial distribution. Let's look at the third choice.

Number of days between solar eclipses. Now, solar eclipses are a very deterministic phenomena. So, it's very unlikely that number of days between two solar eclipses would follow a Poisson distribution. The last choice is number of days until a component fails. This is most likely to be well modeled using a geometric distribution. Suppose, there's a probability 'p' of the component failing on any particular day, independent of any other day. This is precisely the geometric distribution then. Let's consider another example of the arrival process of people at a shopping center.

And, suppose the mean number of people arriving per hour is '18', so lambda is '18', and I want to find the probability that there are '20' arrivals in an hour. Using the formula from the previous slide, this is given by 'e' power minus lambda. Lambda to the '20' divided by '20' factorial. I'm just plugging in the value of lambda, I get this. The main motivation of introducing the Poisson distribution was to introduce the Poisson process. A Poisson process is a counting process that counts the number of occurrences of random event as a function of time. For instance, the number of arrivals as a function of time. So, let me denote accounting process by 'n' of 't', which counts the number of occurrences of that event as a function of time 't'. This counting process or the Poisson process is specified by a single parameter which is the rate lambda, which specifies the mean number of events in a...in a unit of time.

The process 'n' of 't' is the integer valued process, and probability 'N(t)' is equal to 'n' is given by lambda 't' is to the power 'n' divided by 'n' factorial times 'e' power minus lambda 't'. So, the random variable 'N' of 't' is distributed according to a Poisson distribution with rate lambda 't'. Therefore, the expected value of 'N(t)' is going to be lambda 't'. An important property of the Poisson process is that inter-arrival times are independent and are distributed according to an exponential distribution with rate lambda. In other words, if I look at the distribution of inter-arrival times, times between any two arrivals, that is distributed according to an exponential distribution.

Let's see a simple example. We know 'N' of '0' is '0'. And, let's try to find the probability that 'N' of 't' is also '0'. So, there is no arrival until time 't'. Using the formula above, this probability is 'e' power minus lambda 't'. Since there has been no arrival until time 't', the first inter-arrival time must be more than 't'. So, probability 'X_1' is bigger than 't' is equal to 'e' power minus lambda 't'. This implies that 'X_1' is exponentially distributed with rate lambda. We can in fact extend this argument and show that inter-arrival times are independent and exponentially distributed with rate lambda.

Video 14 (3:29): Poisson process: examples

Let's see some examples of a Poisson process. As a first example, let's consider the arrival process of buses at the student center, and suppose the inter-arrival time is exponentially distributed, with mean of 60 minutes. As we discussed on the previous slide, this is precisely a Poisson process with rate lambda being one over '60'. I want to think about the following question. Suppose, a student arrives at the student center at a random point in time, what is the average waiting time for that student? Since inter-arrival times are exponentially distributed and the exponential distribution has the memoryless property, it doesn't matter how much time has elapsed since the last arrival. For a student arriving at a random point in time, the average waiting time is still '60' minutes.
Now, this might appear counter intuitive at first. We know that the average length of the segment between any two arrivals is '60' minutes and if we arrive at a random point in time, we might expect to see our waiting time to be '30' minutes but actually our waiting time is '60' minutes, and this is called the— this is referred to as 'The Waiting Time Paradox'. If you pause and think for a minute, why would you think that arriving at a random point in time would cut your waiting time in half? Since the inter-arrival times are exponentially distributed, arriving at a random point in time, you are more likely to sample a...a time interval which is of a longer length than a time interval of a shorter length, and therefore, the average waiting time is actually '60' minutes. In fact, the easiest way to conclude that the average waiting time is '60' minutes is the memoryless property of the exponential distribution.
It doesn't matter how much time has elapsed since the last arrival, the distribution of time going forward is going to be the same. An important application of the Poisson process is in modeling birth and death process. These arise in many queueing applications. For instance, in the call center application that we discussed earlier. There this counting process would count the number of calls in the system, waiting or being served. In the call center application, there's an arrival process that increases the number in the system, and there's a departure process due to service completion or abandonment that decreases the number in system.
In this mark of chain, we have replaced the probabilities by rates but if you think about transition from zero to one, lambda$_0$ really represents that for every delta unit of time, there's a probability of lambda$_0$delta of going from zero to one. Similarly, we can interpret all the rates as probabilities in continuous time. So, given these properties of the Poisson process, it is very useful as a modeling tool in many different applications, specially queueing applications.

Video 15 (2:18): Lognormal distribution

Another important distribution that I want to discuss is the lognormal distribution. A random variable 'X' is said to have a lognormal distribution if log of 'X' is normally distributed. Suppose, log of 'X' is normally distributed with mean mu and variance sigma square, then 'X' has a lognormal distribution, and the PDF and the CDF can be specified as in the slide here. Both PDF and CDF follow from the PDF and CDF of the normal distribution. Note that the lognormally distributed random variable is a positive random variable.

So 'X' can take only positive values. This is a very important distribution that arises in many natural settings. In fact, it's a consequence of the Center Limit Theorem over log off products of independent variables. So, let's think of a random variable 'Z', which is a product of 'n' independent random variables. 'X_1', 'X_2', so on until 'X_n'. Now we think about log of 'Z', that is going to be sum over 'j' going from one to 'n'. Log of '$X_{j'}$. Since 'X_j's are independent log of 'X_j's are also independent. And, if we think about consequence of the Center Limit Theorem on the sum, the sum is going to be normally distributed. So, whenever we have a random variable, which is a product of independent random variables, we can approximate it well by a lognormally distributed random variable.

Such random variables that are products of independent random variables arise in many natural settings. For instance, in biological settings where we're thinking about growth of a living tissue. In network settings where we are thinking about the spread of epidemic or spread of rumour, number of affected nodes would be Lognormally distributed, for instance. So, the lognormal distribution arises in many natural settings and is a very powerful modeling tool.