[SQUEAKING]

[RUSTLING]

[CLICKING]

PETER KEMPTHORNE: All right. I guess we'll get started. Let's see. I wanted to just finish up some discussion of topics from linear algebra that we didn't get to the last math lecture. And to begin with, I think we did cover this topic of eigenvalues and eigenvectors. These linear algebra concepts turn out to be very, very useful in a number of different applications.

But when we have an eigenvalue and eigenvector pair, then, for a matrix A that has an eigenvector V, then the matrix-vector product is simply a rescaling of that same vector. So matrix multiplication, when we use eigen-- or matrix-vector multiplication when we use eigenvectors is very simple. We're just rescaling the input.

And in order to solve for eigenvalues and eigenvectors, one can solve this equation where the two are equal. And when we solve this equation, we want the roots of the determinant equation to equal 0. This means that these are the eigenvalues which make this matrix factor, A minus I lambda, non-invertible.

So with a matrix, if that matrix has a determinant that is-- or if that matrix is invertible, then the determinant of the matrix is nonzero, and the inverse is actually functions of different determinant cofactors. So in order for v to not be 0, which would be the case if this matrix were invertible, we must have the lambda satisfying this equation.

And this system of equations is a polynomial in lambda, or at least, the determinant equaling 0 corresponds to a polynomial in lambda equaling 0. So these things can be solved very directly. Actually, in R, you can actually compute the roots of any polynomial very easily. And then, once we've solved for our root lambda, then we can solve the equations for what the v vector is that satisfies that.

Now, let me put this in here. It turns out that if we have independent eigenvectors, then we can-- linearly independent eigenvectors-- then we can comprise those as columns of a matrix S. And the matrix A multiplied by this S matrix will simply be the matrix of multiples of the eigenvectors, so S lambda. Lambda is a diagonal matrix.

And the inverse matrix of S will exist so long as we have independent eigenvectors. And so this equation here can be used to obtain A is S lambda S inverse. Alternatively, S inverse A S is equal to lambda. So these eigenvectors, when comprising a matrix, allow us to diagonalize a matrix A.

Now, let's see. All right, so eigenvalues, eigenvectors, can come into play when we have equations for Kalman filters. If our matrix A is transforming states at time t minus 1 into states at time t, and the dynamics of this process proceed regularly over time increments given by t, then we simply have, the t-th transition from state 0 is given by the t-th power of A.

And we can actually get explicit solutions, then, which end up giving us-- let's see. At least, if our eigenvectors, v1 through vn, are linearly independent, then, if we have an initial condition, u0, we can express that initial condition as a linear combination of the eigenvectors. And then we can represent the transition over t periods to have this form. And it ends up just being a sum of multiples of the eigenvectors, with the coefficients being powers of the eigenvalues.

So in looking at this mathematically, one can see that if the eigenvalues, lambda 1 to lambda n, are smaller than 1 in magnitude, then this limiting state transition goes to a zero vector. If one of the eigenvalues is equal to 1, then this term here won't change with each transition, and so we'll have a limiting value of the state that corresponds to a multiple of the first eigenvector.

And so what's interesting to anticipate is that there will be different kinds of problems we work with where the eigenvalues may actually be such that the largest eigenvalue is equal to 1, and the other eigenvalues are smaller than 1 in magnitude. And so we can get nice analytic representations for the transition of the states.

All right. Well, from linear algebra courses, you probably remember that if we have a symmetric real matrix-- so if this matrix here is n by n, and it's symmetric, and it's real, then all of the eigenvalues of this matrix are real as well. And if we have two distinct eigenvalues, then the eigenvectors corresponding to those two eigenvalues are orthogonal to each other.

So if we have A v1 equals lambda 1 v1, and A v2 is equal to lambda 2 v2, then it turns out that v1 dot v2 is equal to 0. And this is useful. At least, if we're-- with linear algebra, we often want to represent multidimensional vectors in terms of some basis. And so, if we think of our matrix A equaling a vector a1, a2, up to a n, if the matrix A has full rank, then these columns of A are linearly independent. And we could use these column vectors as a basis for the n-dimensional space.

So if the rank of A is equal to n, then the collection of aj's, j equaling 1 to n, forms a basis. And if we have a matrix A that maybe is symmetric, but has eigenvalues that are all distinct, then we actually could have-- so if A is symmetric, and all the eigenvalues are distinct, then the collection of eigenvectors, vj, j equaling 1 to n, this would correspond to an orthogonal basis.

And so, when representing n vectors, it's convenient to represent them in a coordinate system that has orthogonal axes in it. Now, what's interesting is that when A is symmetric and real, all the eigenvalues are real. It turns out that there exists an orthogonal set of eigenvectors in that general case. So that can be very useful. All right.

Well, a very important linear algebra result is the singular value decomposition. And let me just ask, how many people here have covered singular value decompositions? Just about everybody, OK. So who would like to explain to me what-- or explain to the class what a singular value decomposition is? So let me write this, maybe, as the singular value decomposition in my notation. So who would like to explain what this is? Anybody? Yes?

AUDIENCE: Expresses the matrix A as a sum of orthogonal [INAUDIBLE] matrices.

PETER KEMPTHORNE: So u is such that u transpose u is equal to, say, the identity that has orthogonal columns. V is the same thing. And D is a diagonal. And so what's important is that any matrix can be expressed as a factor of a V transpose, which is an orthogonal matrix, that will transform coordinate axes to new axes that are orthogonal in the same space.

And then D is a diagonal matrix down to d-- well, I'll write dr-- where we basically are stretching and squeezing axes. And then u simply has columns that define a new basis or a new vector space. And we're taking sums of those. So it's interesting to represent any matrix as this operator, which will do an orthogonal transformation of axes, stretch and squeeze in that transformed space, and then use those to multiply new coordinate vectors, u.

Now, if we take any-- let's say, here, an m by n matrix, A-- and it's often the case that we have maybe m cases in a data set and n variables. And so this corresponds to a data matrix for an analysis of some data. If we take A transpose A, then this is a symmetric real matrix.

And so this actually can be diagonalized to V D squared V transpose, where I'm representing the eigenvalues as the squared D values and the eigenvectors as columns of V. And so one can construct this, do the eigenvalue-eigenvector decomposition here. And then we can take the matrix A, premultiplied by V, and-- or not premultiply-- postmultiply it by V, and postmultiply it by D inverse. And this will basically be our U matrix.

Now, all of this is a bit casual with my notation. The number of nonzero diagonal entries here can be limited to the rank count of the matrix A. And so that's the way it can play out. So we basically have nonzero eigenvalues of A transpose A. There will be r, the rank, nonzero eigenvalues. We can define the orthogonal, orthonormal eigenvectors corresponding to those.

And then, if we want to expand the orthonormal basis of the first r eigenvectors to be an n-dimensional basis, although they essentially aren't used, and so that gives us our singular value decomposition. And it turns out that in dealing with data in a matrix, the singular value decomposition allows us to reduce dimensionality. Perhaps there's variation in the data values that is limited to a subspace of rank 2 or 3. And so we can identify and specify that subspace easily with the singular value decompositions.

All right. Let's see. So in these notes, I refer to reduced SVD. What's rather interesting as a decomposition of the matrix A is that it's a sum of r terms, where r is the rank of the matrix. And those r terms are the products of the column vectors of u times the column vector transpose of v, respectively, with the singular value, sigma 1 through sigma r. So it provides a nice way of thinking of a matrix. It's simply a sum of rank 1 terms.

OK. Let's see. These lecture notes on linear algebra finishes with the Perron-Frobenius theorem. And this theorem relates to matrices that are square with strictly positive elements. And so, with such matrices, the theorem states that there is an eigenvalue, a real eigenvalue, lambda 0, that is bigger in magnitude than any other eigenvalues.

And corresponding to this eigenvalue, lambda 0, we can define an eigenvector v, which has elements that are all positive as well. And this lambda 0 eigenvalue, the largest one, is alone in terms of, all other eigenvalues are smaller in magnitude. And this result turns out to be very useful in working with positive matrices.

The notes here go through the proof of that. I guess they're included here, if we wanted to get into technical proofs of linear algebra. And I don't want to do that. Let's see. I finished these notes with just some mention of different news sources that you should consider exploring in finance. Well, there's Yahoo Finance, which is where we can easily collect price data.

But you should know that the Wall Street Journal is-- you can get a free subscription with that through your Kerberos password-- or ID, rather. So that's very useful. And then there are a number of other sites that you may find useful.

Now, let's see. In our very first lecture, I indicated how we could use RStudio to collect data, plot time series of different stocks or other assets. And in this lecture 2, I distributed this workspace, which has R scripts and output for looking at portfolios of stocks and considering equal-weighted portfolios with different components in the equal weights.

And so if you go to the Rstudio Cloud project-- so with RStudio Cloud, or maybe it's Cloud Rstudio, one can upload different R programs in this project. And so has anyone in the class yet used RStudio Cloud for any of the examples? I'd like everyone to at least be capable of doing that. And I'd like proof that it actually works with students doing it. So it would be great for some of you to consider that.

What's nice about this environment is that I can give you an R program. And it should run just by saying Run. You don't have to install anything. And with this output, there was one program called Script_SP500. And there's a table of contents for this R script, which says it loads libraries, then collects stock prices, and then saves the workspace. So the commands are all very simple in here.

And then, with the Equal Weighted Portfolios script, it actually loads libraries and loads the data set of stock prices that was just downloaded in the other program. And it loads the data in. Then, in processing the data, I-- consider investing a portfolio by equally weighting allocations to different stocks in the S&P 500 index.

And so, if one does this, these simple R commands create the value of such a portfolio, where, in fact, as indicated here, it's investing $1,000 at the beginning of 2019. And as an alternative to just equally weighting across the S&P 500 stocks, I looked at the Amazon, Apple, NVIDIA-- or no, Netflix and Google. These were actually the really hot stocks about four or five years ago. And basically, look at equal weighting of all these. And so here's how the investment, equal weighted at the beginning of the period, resulted towards the end.

So there are some interesting features one can see. If one happened to be good at identifying stocks to invest in, you can do much better than the average S&P 500. There can be greater risk by investing in a smaller number of stocks because you're not so diversified. And so you can see how the red graph of values decreases rather dramatically at some points in the time period.

And if one looks at what happened to the $250 invested in each of those four stocks, this is what happened over this, whatever it is, five-year period, or almost five-year period. And this raises some interesting issues for us, which is, when considering portfolio construction, maybe equally weighting allocations across some assets is a good thing to start with.

But what happens over time? Well, in this case, Apple outgrew the other allocations substantially. And so the degree of diversification in the portfolio is getting lower because we have a greater concentration in that one stock. So what we'd like to do, perhaps, is to rebalance the portfolio. And different approaches to rebalancing portfolios can be considered.

Now, if we were to consider rebalancing every day the investments in these four stocks, then would you think that's a really good way to rebalance, to do it very, very frequently? And why or why not? Yes?

AUDIENCE: Well, if you do it too frequently, then variances like 2022 can affect it. It won't be very good for the big picture.

PETER KEMPTHORNE: Yeah. If you rebalance too quickly, then you're taking funds away from winners and giving those to losers. And so you're equalizing allocations in a way that doesn't benefit from short-term trends or strengths in different assets. So that extreme is probably not something you want to do.

One thing that's rather interesting is just how, if you do not do any rebalancing, then the portfolios become rather undiversified. And so the risk level is actually probably increasing over time here. So one way of rebalancing portfolios is trying to maintain a certain risk level for the portfolio. And such an objective would require periodic rebalancing.

OK. Well, let's move on, then, to the topic of probability theory. And so what I want to do in this note is to just give you a quick overview-- basically, a review of probability concepts that you probably are very familiar with.

And so, with probability theory, it's customary to distinguish between discrete-valued random variables and continuous ones. And so, with discrete-valued random variables, those can be just events occurring or not, like a default of a counterparty, or maybe the Federal Open Market Committee deciding on a Fed funds rate.

In thinking of individual stocks and their dynamics intraday, one can think of different orders for a stock to buy or sell coming in at random and wondering what side those would be on-- also, perhaps, the sheer size of the next market order. Now, it used to be that these were always discrete outcomes. Like, share size, it turns out you can actually now invest fractional shares in different stocks. So the market's become more rich in that way.

But, importantly, we do want to distinguish between discrete outcomes that have probabilities with respective outcomes, versus continuous random variables, where we're thinking of values that could take on any value in some interval or range. And so, in thinking about the values of assets, we can think of the value being in a continuous range.

Also, timing of events is an important random variable that is continuous as well. So one might consider, what is the waiting time to the next market order for Apple stock? Mixed variables can arise where we think of a stock having some value, where it's generally a positive value, but it could go bankrupt, in which case there would perhaps be some discrete probability of bankruptcy for a stock.

Now, with probability models, we use probability mass functions for discrete variables and probability densities for continuous. And so we use calculus to calculate continuous probabilities in different intervals. And, importantly, the cumulative distribution function is a very important tool.

So if we write F sub x of-- let me write c-- is equal to the probability that the random variable x is less than or equal to c, as a function of c, this value is going to be no smaller than 0 and no larger than 1. And it will be monotone increasing in c. And distributions will be uniquely specified by specifying their cumulative distribution function.

So these are very simple concepts. There are expectations in moments. One just-- I'll just call this an anecdote-- if x is continuous random variable, and we define y equal to f of x, so we consider the probability integral transform of the random variable x, what is the distribution of y? OK.

AUDIENCE: Uniform.

PETER KEMPTHORNE: Yeah, it's uniform.

AUDIENCE: On 0. It's uniform on 0.

PETER KEMPTHORNE: So what's really neat is that a monotone transform of the random variable x, whatever x is, so long as we know what the distribution is, transforms that to a uniform. And so this is true regardless of what F is, so long as F is continuous. And that property can turn out to be very, very useful.

One issue could be, oh, suppose you have a random sample of data from some distribution, or you think it's from some distribution. Is the sample consistent with that distribution or not? Well, if you did this probability integral transform, it should lead to a uniform distribution, which may or may not be realized. So anyway, that turns out to be quite useful.

Now, we have expectations and moments of random variables. Basically, the mean of random variables, the probability-weighted outcome, we can generalize the mean to looking at, which is the first moment, to consider the probability-weighted average of the k-th power, where k is an integer.

And the variance of random variables is a useful measure of uncertainty about the mean value. And there's a nice formula for the variance, which is the squared distance from the mean in terms of the average squared minus the square of the average. So these are useful formulas that may be useful at certain times.

Now in terms of measuring uncertainty about outcomes, instead of the variance, taking the square root of the variance is useful because its scale is the same as the original units. Most people don't think in terms of squared units with any problem. And so, instead of dealing with the variance directly, we'll commonly take the square root of that to have the same units as our original variable, x.

Now, with the skewness and kurtosis, we're looking at measures of the shape of the distribution. And with skewness, as you might anticipate, that detects asymmetry in a distribution. And so, something that's nice to think about is, if we have a random variable x, where the expected value of x is equal to mu, and the square root of the variance is sigma, if we take z equal to x minus mu over sigma, this is a standardization of the random variable.

The expected value of z will equal 0. And the variance of z will equal 1, as will the square root of the variance. And for the skewness, which I use-- I guess it's a gamma-- gamma is equal to the expected value of z cubed. And the kurtosis is equal to the expected value of z to the 4th.

So the skewness and kurtosis parameters of a distribution can measure asymmetry. At least, if the cubed value has more weight on the positive end than the negative end, it'll be skewed to the right. And with the kappa, we're looking at z to the 4th. So we're looking at, basically, how heavy are the tails?

And for a Gaussian distribution, kappa is equal to 3 for a normal, 0, 1, or any normal. And gamma is equal to 0. So anyway, so these other parameters are useful for characterizing distributions that are not Gaussian-- non-Gaussian in terms of, perhaps, skewness or heavier tails.

All right. Well, let's just review Gaussian distributions. I hope this formula for the density of a Gaussian is very familiar to everyone here. We basically have, the density function is equal to 1 over root 2 pi sigma e to the minus 1/2 x minus mu over sigma squared.

And so the mean is mu. The variance is sigma squared. And the density function is symmetrical, with a maximum at the mean, mu. And the points of inflection are actually at plus and minus sigma. So if we think of just trying to draw a normal density function, then the peak is at the mean. And the point of inflection where the second derivative is 0, it goes from negative second derivative to a positive second derivative. That distance is sigma, symmetrically.

All right. So let's take a look at the normal bell curve. 68% lies within 1 standard deviation. So this is whatever-- AP Statistics, you learn these facts. If we go to 2 standard deviations from the mean, that's 95%. And 3 standard deviations is 99.7%. So these models are-- normal models have these properties of how likely different variations are in standard deviation units.

Well, in finance, we'd like to exploit, perhaps, Gaussian distributions for modeling prices. And with a normal density curve, it actually applies to x values that can be arbitrarily large, negative or positive. It's an unbounded distribution in terms of potential outcomes.

And so one way of dealing with that with, say, prices is to consider taking the log of prices and thinking of that being a normal distribution. And so that construction corresponds to a lognormal distribution. And so if we have a random variable x that's normal, and then we take the exponential of that normal, then we say that the y variable is lognormal, with the parameters mu and sigma squared corresponding to the log scale.

Now, what will be very interesting in subsequent lectures is when we talk about stochastic processes that can be applied in financial markets. There's an important stochastic process called Brownian motion, which many of you probably have some familiarity with. And with a Brownian motion process, the increments of the process over time periods has a distribution that's normally distributed.

And as the length of the increment of time increases, both the mean and variance of the outcome changes. And so this Brownian motion process is a very, very interesting process that is intimately related to Gaussian distribution. So we'll get there in a few lectures.

Now, one of the useful tools we have in probability theory is being able to derive the distributions of transformations of random variables. So with the lognormal, we have a normal distribution. We maybe take the exponential of that normal outcome. And we'd like to be able to write down the density of that distribution and compute different features.

Well, we can think of, say, a random variable x being transformed to y with a function g. And then we can consider a function h, which is the inverse function of g. And as examples, if g is e to the x, then h is the log function.

And with change of variables, if we have the CDF of x, then the CDF of y is equal to the CDF of x evaluated at the inverse of y on the x scale. So this equation here is a simple logical statement that everyone here, I think, would agree is valid.

And so, if this function relating the distribution functions is true, if we take the derivative of the CDF function of y, we get the density. And then we just have to apply the chain rule on the right-hand side, the density of x evaluated at h of y. But then we also need to have the chain rule for h in there.

So with this formula, it's all familiar. But if we have f of x is equal to the integral minus infinity to x of f of t dt, then the derivative of f evaluated at x turns out to be equal to f of t evaluated at t equal to x. So the cumulative integral has a derivative equal to the argument of the integral.

So this is easy. And we can use this to get the density function for a lognormal distribution. And here's the density of the lognormal distribution. It has the same form as the Gaussian density, plugging in log y for x. But then we need to multiply that by the chain rules, or the derivative of h, with respect to y, which is 1 over y.

So here's the lognormal distribution. And as an example, I chose a mean log value of 0.2 and a SD on the log scale of 0.4. And as you work with different distributions, you may have-- not suggestions, but you may find it useful to consider specific parameter values.

Here, I'm thinking that if one has a-- if we think of, say, x of 0 equaling 1 and x of 1 being distributed as a lognormal, with mu equal to 0.2 and sigma equal to 0.4, then this lognormal distribution can correspond to a one-year period where, on the log scale, we expect a mean change on the log scale to be 0.2, and with standard deviation of 0.4.

So this would potentially be relevant for an asset and looking at its percentage change on the log scale over 1 year, with an annualized return of about 20% and annualized standard deviation of 0.4. Well, if we draw vertical lines in this distribution space, or sample space, of the lognormal-- here, I've just drawn symmetrical percentiles-- 25th and 75th percentiles, and then the 5th and 95th percentiles in red. And it's obvious that it's asymmetrical.

And what's useful, actually, with different distributions is being able to calculate quantities of interest in finance. And so here's a nice example of a financial computation, where we consider a call option to buy an asset at some terminal time, capital T. And the strike price is going to be at price K.

And so if we think of X being the random variable of what the time T price is, then the payoff is going to be equal to X minus K, the value minus the strike price, if that's positive, or it won't be worth anything. So if we have, at time T-- if we were to draw here as a function, P sub T-- or sorry, X, I guess-- and there's a strike price here, then we would have the 0 value up to K. And then we'd have X minus K be giving this option payoff.

Now, you'll hear Vasiliy and Jake talk about the hockey stick payoff function of options. And so this is what that comes from. Now, what's interesting to know is this theorem here says that if we have a density function little f, with CDF capital F, for the outcome at time capital T, then the expected value of this payoff is equal to this integral from the strike to infinity of 1 minus F of x.

And the proof of this is done by considering a finite upper limit, M, capital M, and then integrating by parts and getting a formula here. So it's an interesting result that if we know the terminal time T random price of an asset, then a call option on that terminal value will have this expected value.

So those of you who actually have some advanced probability coursework may notice that when you are looking at the expected value of a positive random variable, you can sometimes express that as an integral of the complementary probability of exceeding that value. And so that result ends up being applicable here in a very nice way.

Now, with a normal distribution for X, we get this formula coming out for that call option. And if we have a lognormal, we get this different result. So knowing the distribution affects what the option price is.

Now, in probability theory, we make use of moment-generating functions, which, I guess, when I first learned of these, I thought, well, nice tool, but I don't really care. At least, it's a highly technical tool to use. However, it turns out these are incredibly useful, moment-generating functions.

And the definition of the moment-generating function is given by looking at just the expected value of e to the tx. And we can expand this argument in a Taylor series, which is done on the first line there, and just take the expectation.

Now, something that we should be questioning, perhaps, is, do moment-generating functions always exist? And by that, I mean, do we get a finite-valued function, mx of t, for all random variables? OK.

AUDIENCE: No? Sorry.

AUDIENCE: No, that's fine. No, that's fine. That's fine.

AUDIENCE: Yeah. No, no, no, that's-- yeah.

PETER KEMPTHORNE: The moment-generating function doesn't have to exist. And the moment-generating function won't exist if this exponential function has too much weight. So that integral is infinite. And an example where it is infinite is a Cauchy distribution. A Cauchy distribution doesn't have finite mean or variance.

But what's interesting is that-- this is called the moment-generating function. There's a characteristic function-- characteristic function-- that turns out to be equal to e to the itx, where i here is the square root of negative 1.

And if you know a little bit about complex numbers, this turns out to be equal to the cosine of tx plus i times the sine of tx. And it turns out that these characteristic functions always exist and can be used almost the same way a moment-generating function can be used when it exists. So the fact that it doesn't always exist shouldn't worry us because we can generalize to characteristic functions.

But what's important is that if we have two random variables that have identical moment-generating functions, then those distributions are equal for the two random variables. So we have identical distributions if their moment-generating functions are equal.

And if we consider a sequence of random variables, and the moment-generating functions that can be thought of as a sequence converges to a function that's a moment-generating function of some distribution, then we have that the sequence of random variables, Xs, converges to that distribution as well. So these are nice properties that allow us to understand limits of random variables and what the distributions are of limits.

Now, with moment-generating functions, this slide here discusses-- well, it introduces you to the moment-generating function for the normal distribution with mean mu and variance sigma squared. And with the special case of mu equals 0 and sigma equaling 1-- for a normal, 0, 1, if we look at the expected value of e to the tx, it's equal to the integral of e to the tx times e to the minus 1/2 x squared over root 2 pi dx.

So this is simply the formula for e to the tz, I guess. Well, if we look at this exponential here, we can write this as e to the minus 1/2 x minus-- let's see-- x minus t over 2 squared. That gives us x squared minus-- oh, no, just x minus t squared.

Is this right? I basically want to take 1/2 x squared plus tx and write this as minus 1/2 of x minus t squared minus t squared over 2. Is that right? I want a plus right there.

So this exponent of the exponential here can be rewritten as this exponent times that. So this is equal to the integral from minus infinity to infinity, 1 over root 2 pi, e to the minus 1/2, x minus t squared, and then plus t squared over 2 dx.

And so this exponential factor with t squared can be taken out. And this integral here is simply going to equal 1. So we are able to find the moment-generating function for the standard normal very easily. It's equal to e to the 1/2 t squared.

Now, if we think about transforming a random variable, such as X, with a linear function-- basically, rescale X by sigma and add a mu-- then the moment-generating function of that transformation of X to Y can be written out. And it's equal to the exponential of t mu times the moment-generating function of X evaluated at t sigma instead.

So knowing what the X moment-generating function is, we basically get this function coming out for the moment-generating function of a normal. Now, this moment-generating function is, in fact, the expected value of a lognormal random variable.

At least, if we set t equal to 1, it's the expected value of a lognormal. And if we take t equal to 2, it's equal to the square of a lognormal distribution. So we can use the moment-generating function for the normal to calculate moments of a lognormal distribution. And that'll be an exercise in the next homework.

OK. All right, so linear transformations-- basically, linear transformations don't change the skewness of distributions. And the kurtosis can-- well, it might affect the kurtosis. All right. Well, let's move on to considerations of more than one random variable and the concept of two random variables being independent.

In probability theory, we say two random variables are independent if the joint probability of the random variables, X lying in a set A and Y lying in a set B, if that joint probability of both occurring is equal to just the product of each occurring alone. And so we have, for densities or probability mass functions, we have this property of the joint density function factoring into the product of the marginal densities of each of the random variables. And it works for a continuous or discrete distributions as well.

So independence means that, basically, knowing one variable doesn't give us any insight on the value of the other variable. Now, we can compute covariances between two random variables and correlations. And so, with a covariance between two random variables, this is a generalization of the variance for one variable. But this is the probability-weighted average of x minus-- I'll write here mu x times y minus mu y.

And the correlation statistic corresponds to dividing this covariance by the product of sigma x and sigma y, the square root of the variances. And it turns out that this also can be written as x minus mu x over sigma x times y minus mu y over sigma y.

And so the standardization of the random variables to have mean 0 and standard deviation 1 can be made. And then the expected product of those standardized values is the correlation statistic, or the correlation parameter.

Importantly, zero correlation doesn't imply independence. But zero correlation does mean that the dependence-- does suggest that there's no linear dependence between the two variables. Now, when we consider generalizing from two random variables to many, then we can think of working with random vectors. And with, say, little n being the count of different random variables, we can have notation of mu j for the means or expectations of each, and sigma i, j corresponding to the covariance between Xi and Xj for sigma i, j.

And we can then represent-- the expected vector of x values is equal to a mu vector. And we can represent the covariances between the elements in terms of an n-by-n matrix, sigma. And so what's important to realize here is that if we have an x vector and a mu vector, so x is equal to x1 down to xn, and mu is equal to mu 1 down to mu n, then we define the covariance matrix of x to be the expected value of x minus the mu vector times x minus the mu vector transpose.

And so this is an n-by-1. This is a 1-by-n transpose. And so we get an n-by-n matrix, where we take expectations element by element in that array. So this is a nice way of representing the covariance matrix.

And what's useful in particular is, well, if we think of Xs maybe representing returns on different assets, with means mu, and the covariance matrix tells us how much variability there is in each of the Xi's, but also the correlation structure across the n Xs, then if we consider, say, a linear combination of the Xs given by a vector A of constants, then this linear combination of the Xs can actually represent maybe a portfolio of holdings across different assets given by the Xi's.

And what we'd like to be able to do is model the mean value of Y and its variance. The mean value is very simply the linear combination of the component means times the a coefficients. And when we go to calculate the variance of Y, if you go through each step here, each equation here is doing something slightly important.

We're basically substituting in A transpose X for Y at the beginning and a transpose mu for the expected value. Then we are able to pull out the constant vector, a transpose, in the front of the expectation, and actually in the back of the expectation. So these coefficient vectors are constants and can be pulled out of the expectation computations.

And then this middle term here is simply the covariance matrix of X. So we have that the variance of Y is simply the sum, double sum, over all pairs of coefficients, ai, aj, multiplied by the covariance terms between i and j. So this is the formula for the variance of Y.

I'm not sure if we go into this in the next slide. But if the covariance of Xi and Xj is equal to 0-- so if this is equal to 0 for all i not equal to j, then this variance expression simplifies a lot. And if, say, aij-- sorry, ai is equal to aj is equal to 1 over n, then the sum of the ai's is equal to 1. And the sum of the ai squareds is equal to what?

AUDIENCE: 1?

PETER KEMPTHORNE: No. No, this is not equal to 1. It's going to be less than 1.

AUDIENCE: 1 over n.

PETER KEMPTHORNE: 1 over n, right. It's n, n times 1 over n. So it's-- sorry, n times 1 over n squared. So the variance of an equal weighted average of the Xs will have variance that's 1 over n times the variance of each of the Xs.

So this is where diversification can be useful in terms of reducing variability. And in the very special case where we have, say, zero correlated assets and, I guess, equal weighted variances of the Xs, then we can reduce the variability by the factor n, or multiplying by the factor 1 over n.

Well, this leads us to introducing you to principal components analysis. And with principal components analysis, we consider an m-variate vector x-- so we consider an x vector that's m by 1, which has an expected value equal to an alpha vector, and a covariance matrix, which is m by m, equal to sigma matrix.

And this covariance matrix is a real matrix that is symmetric. Therefore, it's going to have real eigenvalues. And it's also going to have nonnegative eigenvalues because the covariance matrix is positive semidefinite. And so sigma is positive semidefinite.

The definition of positive semidefinite is that for all a vectors, a transpose sigma a is greater than or equal to 0. And this is actually the variance of a transpose x, this product of the covariance matrix times the transpose of an a vector and that.

So the fact that the variance of the linear combination of the x's is strictly nonnegative gives us this result here, which means that the sigma matrix is positive semidefinite. And we can basically list the eigenvalues to be lambda 1 through lambda m. Those could be all different. Some of them will be nonzero, so long as there's no-- so long as the covariance matrix has some positive rank.

And so, with this framework, where we have the covariance matrix, sigma, having eigenvalues, lambda i, and corresponding eigenvectors, gamma sub i, we can normalize the eigenvectors, gamma i, to be of unit length. So gamma i transpose gamma i is equal to 1. And we can also define the gammas to be orthogonal to each other.

Then principal components defines new variables by saying, we take our vector x. We transpose it to x minus alpha. So we're looking at deviations from the mean values, component by component. And then we take each of those, and we multiply them, or look at the dot product of gamma i dot x minus alpha.

And we call this the coefficient of the i-th principal component variables. And these principal component variables are very neat because their average or mean value is equal to 0. But the covariance matrix between the p's is--

[ALARM RINGING]

It's time to finish the lecture here. The covariance matrix is the diagonal matrix lambda. So we basically have transformed our original x vector into new coordinates, which have mean 0, are uncorrelated with each other, and have variability that decreases, basically, with i, if we rank the principal components, eigenvectors, lambda 1, to be bigger than-- or basically, to be ordered from lambda 1 to lambda m.

So we have an orthogonal basis for our data in this case. And I guess we'll finish there for today. But on Thursday, we'll finish these notes up. And what you'll be able to see is how we can try to represent an x vector as a sum of terms. And maybe the sum of terms corresponds to using just a small number of the principal component variables.

So when we have multidimensional vectors that have variability that is limited to a smaller number of dimensions, we can characterize those smaller number of dimensions nicely with principal components and obtain useful models. And these kinds of models underlie multifactor models often in modeling asset returns and financial markets. So we'll finish there for today, and thank you.