[00:01:00] Welcome to the Wednesday, September 1st, webinar.
Today the plan is to first go through quiz 1, or at least the big.
I'll give you some hints and thoughts about that.
It's due next Monday, which will give you a chance to decide whether to stay in this course or drop it before the census date.
Then the next item is that we will answer the questions posted to the thread that have not been answered in the Q&A thread.
I think what I will do is I know what Sarah Ray requested that I derive the multivariate normal density.
That's a little bit outside the scope of this course, so I think what I will do is I will leave it to the end.
The others I will do in order posted.
[00:02:00] Any questions before we start? First of all, there was some confusion in the quiz and I apologise for that.
Basically, the description said it was a 90-minute time limit.
That was not correct.
I updated it everywhere except for that one place.
I apologise for that.
There was no time limit to this quiz.
You have all week to work on it.
You can stop at any point and then come back to it later.
Now I know that one person, actually, I think under the assumption it was 90 minutes long, rushed through the quiz and submitted it.
Again, I think I've [00:03:00] reopened the quiz for them, but if you also have that problem, please let me know.
But again, there is no time limit on this week's quiz.
I haven't decided how do the others I need to go see how this one goes first.
But I may or may not have time limits on the others.
Now, I will also like you to submit your working for this quiz into Turnitin.
This is necessary both for purposes of academic integrity, because unfortunately, while and obviously everybody here is wonderful and perfectly honest, we have had quite a bit of collusion and plagiarism problems.
We're trying to address that, but the other benefit of those is that if for example you have submitted a question and [00:04:00] you say mistype something in the answer, and the answer blocks the quiz, but if it's correct or mostly correcting your working, you can send me after you had your mark saying, hey Paul, maybe I deserve some extra marks here, I may correct.
I'll look through and I will see if you are entitled to extra marks.
By the way, sorry, not an email form.
I actually prefer that you post to the forum, interact through the forum, it keeps things a bit more organised.
The quiz, you can use R or any computer algebra systems, they are permitted and even encouraged, but as part of your working, you should submit the code you used to get your results.
There are certain technical issues [00:05:00] and model where if you leave a question blank, it might not be able to mark the rest of the questions correctly.
It's rather annoying, but if that's an issue, we can work around it, but it's probably easier to just if you filled in all your answers if you don't absolutely have no idea, well, first of all, you should ask for help, for hints and such come to a consultation.
But if you absolutely have no idea and haven't been able to answer it, then please put in zero or some other numerical placeholder.
Now, there was a question of level of significance.

Actually, this refers to significant digits.
I recommend not rounding intermediate results at all.
Pre-calculations, and if you're using [00:06:00] R, that means instead of you assign the output to a variable and then you reuse the variable.
But for final answer is usually four decimal places suffice.
There is four places after the decimal point, but you won't lose marks for being more precise atleast for the quizzes.
All right.
Now, this page is basically a revision of your multivariate probability skills.
You basically have to derive some marginal distributions, conditional distributions and then some expected values and variances.
Here the question, what is the most integration? Question 2, you can answer a number of ways, you do need to figure out the characteristic polynomial.
The form of the polynomial [00:07:00] is in the lab notes and in the lecture notes.
The eigenvalues and eigenvectors, actually R will give that to you, and similarly for the Sigma to the one-half and semi-inverse that's discussed in lecture notes.
I think the software demo and challenges have some hints for that as well.
This one is about confidence regions or t-squared confidence regions and chi-squared confidence regions, because that's a hint for one of these parts.
Generally, I suggest, again, looking at the formulas and think about what they represent geometrically.
Again, [00:08:00] the demonstration code actually has the ingredients for all these calculations.
Now the simultaneous confidence intervals, I will talk a bit more about those because somebody else asked a question about that.
All right.
Next set of exercises, this is meant to, first of all, how do you conduct a t-square root test and also do some normal distribution diagnostics.
Even here you have a dataset with a total of five variables.
Of these the river which I tried to spend two years in Pittsburgh, so I like this question.
This dataset in particular, but they're Pittsburgh is a city built at the crossing of three great rivers Monongahela, [00:09:00] Allegheny, and Ohio.
Ohio is also a state, but it is also a river.
Now, the thing is, somebody basically compiled a dataset database of the bridges that had been built over these rivers that vary in various places, and I've taken an extract of that dataset with the which rivers crossed, when it was built, so you can think of this as also computing the age of it.
The purpose of the bridge, the type of the bridge, that's categorical.
Length in feet, that's quantitative and the material used for the bridge.
We have a bunch of categorical and quantitative variables.
Here we will focus on quantitative ones, which is the year it was built, and the length in feet.
By the way, can you see them? Yes, you can see the last cursor, good.
Just checking.
You can click here to download the dataset.
Here I want [00:10:00] you to do some diagnostics and perform the normality tests that were shown in the software demo and the notes.
Basically, I want you to tell me which of these are consistent with multivariate normality or univariate normality.
Then I want you to regardless of that, let's assume that you have normality, and I want you to perform a two-sample t-squared test.
This is just to report the numbers from the test, and then state the conclusion at the conventional 0.

05 level.
Now, here you need to think carefully about hypothesis test and what does rejecting null hypothesis or not rejecting null hypothesis actually [00:11:00] means.
Then there was a series of multiple choice questions, true or false questions.
There is a guessing penalty in effect, which means that random guessing gets you some zero marks.
But if you have even the slightest clue about which answer is correct then just answer it anyway.
So one is a true or false question, and then there's a series of questions of the form.
We've learned in your notes, you'll find some statements and properties like these.
Some of the questions here is whether these apply always to all distributions for which can make sense, or do they only apply to the normal distribution and possibly others? It might apply to distributions other than normal, but not all of them.
[00:12:00] The answers for all these will generally be found in the notes.
Those are the questions and the themes of this quiz.
Any questions? By the way, since there are only 11 of us, I wonder if you all want to turn on the cameras and maybe just have a bit of more of a dialogue and unmute yourselves.
[00:13:00] First, it was a question by Tron and again, if I'm mispronouncing your name, please let me know.
If you only have one attempt can come back later? Then the answer is yes.
Well, so the idea is that right now you might answer a bunch of questions.
Here and here and so on, but then until you click Finish Attempt, your quiz is not submitted, you can navigate away to another page where you're going to close this page and come back to a bit later and the answers will still be there.
But it's not until you click Finish Attempt that your quiz is actually submitted.
You don't have to do it in one sitting, but you do only get to click Finish Attempt once.
That said, if you did accidentally click Finish Attempt, and then realised that [00:14:00] you need to do something differently, let me know when I should be able to unlock the quiz for you so you can update it provided it's before the due date.
So next question by Shane.
I'm confused by question 1 c and d.
Let me pull that up.
These quantities are in the notes.
I think maybe I'll give you a chance to look at that and then I might post a hint at some point.
But it is in the notes, the question prediction [00:15:00] and the properties of multivariate normal with respect to that, but also other distributions.
I think I'll hold off on that and post a hint later.
On further thought let me just point you to the page which has the hint.
You [00:16:00] can change that.
There it is.
There you go.
The link is in the chat, so take a look at that part and that's my hint.
Any other questions about the quiz? [00:17:00] Moving on.
I need to switch which window I'm showing here so one second.
That takes more clicks than it should.
Was there mental equality in the end of Q1c? One second.
[00:18:00] Actually, I think there's a typo there.
I think it should be less than infinity.
Good catch.
I will fix that.

Apologies, there was some kind of formatting error there.
Thank you Tai.
The next item was the request from Jay to talk about properties 4 and 5 of the normal distribution and their applications.
[00:19:00] Let's do that.
Again, I'm now sharing the right part of the screen.
These are the properties of multivariate normal distribution.
We're not going to prove them unless you want me to prove them, in which case I will see if we have time.
Properties 4 and 5.
Property 4 is about prediction fundamentally.
The idea is that you have some variables that are all jointly multivariate, normal.
We split them into two groups.
Let's say that the group 1 is the variables we're predicting and group 2 so X_2 is the set of variables that we're predicting from.
Makes sense? [00:20:00] We have those variables.
They have their means, they have their variances and they have their covariance, the covariance between the first group and the second group.
Let's also assume that these variables that we're predicting from, well, the covariance matrix is full rank.
I think somebody else asked to clarify that and I will get back to it later.
In fact, this means it's invertible.
Now, let's think about given that we know X_2, what can we say about X_1? It's distribution.
It turns out that what we can say about it is almost like linear regression, we can say that if we know X_2, then the expected value of X_1 [00:21:00] is going to be, well, first we start with the baseline expected value of X_1.
Then we have this part.
Now, let's think about what it is.
Well, this bit that is X_2 minus Mu_2, that's the residual.
That's how much bigger X_2 is than its mean.
Then this is essentially the inverse variance of X_2.
We're taking X_2, X_2 is residual, subtracting the mean and dividing by the variance, not by standard deviation but that might seem a bit familiar.
It's almost like a z score except [inaudible 00:21:54] one more time.
Then we multiply [00:22:00] back by the covariance of X_1 and X_2.
In other words, this is essentially the strength of a linear relationship between X_1 and X_2.
Actually if you look at the univariate settings of Sigma_12 is just the covariance between X_1 and X_2 and Sigma_22 is a variance of X_2.
That would actually be the slope of the regression of X_1 and X_2.
You get something.
Now, also, it's worth observing that these Sigmas are just constants.
This is going to be a constant matrix.
It will be some Mu_1, some intercept plus [00:23:00] some matrix which is a slope of our univariate, and then times the residual.
Another thing we can observe its regression like is that if X_2 equals to Mu_2, then our prediction for X_1 is Mu_1.
That also makes sense because in something like a regression, the linear regression line passes through the point x-bar, y-bar.
That's the mean.

The variance.
Well, we started out with Mu_1, Mu and Sigma_11, that's the starting variance of X_1.
Then we subtract off [00:24:00] this quantity.
This is going to depend on a few things, but one thing it certainly depends on, which is very important is Sigma_12.
Now, if Sigma_12 is relatively big, that means we're going to subtract a pretty big number from Sigma_11.
That means that if Sigma_12 is relatively big, at least compared to Sigma_22, that means that X_2 explains a lot of the variation in X_1.
That means that after we take out the effects of X_2, we'll have less variation left in X_1.
By the way, one simple observation is if Sigma_12 is just a matrix of zeros, then X_2 is just not going to do anything to predict X_1 and Sigma_11 will be the variance of the prediction.
If the two [00:25:00] vectors are not correlated, then that's what it is.
Then they have no predictive power for each other.
That's what this property is about.
Makes sense? Now, there was in fact a univariate example here, which I referred to a few times.
But the basic idea is that this is actually the slope of regressing X_1 on X_2 and so on.
This you can view, again, in the context of linear regression as the residual variance.
Next is property 5.
[00:26:00] Property 5, it's going to be important for things like hypothesis testing.
What does it say? Well, it says that if you have some normal distribution that's p variate and has mean Mu and variance matrix Sigma.
Then if we write down this expression in terms of X, Mu and Sigma, which you can think of this as if this were univariate.
This would be X minus Mu, so we get X minus Mu squared and then you divide by Sigma, which is the variance.
It's X minus Mu squared over Sigma squared, which is the z-score of the distribution squared.
We know that whatever these quantities are or whatever these Mu's and Sigma [00:27:00] are.
If X has a distribution specified here, then this is going to have a distribution of chi-squared with p degrees of freedom.
One reason why this is very important is that it essentially lets us summarise how some observation deviates from some hypothetical mean even though they're multivariate.
Let us summarise it in one number and it tells us how this number is distributed.
That's where this.
Now, we'll be seeing it again and again in the context of manipulations and tests based on multivariate normal distribution.
Now, there was also this remark [00:28:00] here, which is that where does this come from and the idea is that if you've seen chi-square distribution before, you get that distribution if you take p standard normal random variables are independent of each other and then you square them and then you sum them up.
The result is going to be chi-squared with three degrees of freedom.
This bit here is basically showing why does this expression always hold? Because of the way you can write this expression as Y transpose Y, where you can show that each of the Y's [00:29:00] is half of this product.
Well, that it's this part is split in half.
You can show these Y's are distributed normal with mean zero and variance one for all variables.
Now, there are some remarks about prediction.
Again, I would suggest taking a look at those.

They might be useful on the quiz.
If you still have questions, I guess put it in the next webinar thread.
Any questions about this bit? [00:30:00] Sarah.
Next slide.
Sorry, Sarah, an example of what? Where you can work out means where you can use these properties to make a prediction or do you mean where we derive these properties? I don't have an example off the top my head.
Let's put a pin in that and I will [00:31:00] perform together for the next webinar.
I think there might be one or two exercises.
I have been talking about the applications and general terms, how this will come in in the future.
But again, part of the problem is here we're laying the groundwork.
For example, this property and in particular, this property is absolutely critical and we're going to be dealing with canonical correlations, but we won't be dealing with them until Week 3.
A lot of this is setting up for the rest of the course.
But I will try to come up with something.
[00:32:00] I think the next question was from Elisa Simpson to discuss the general multiple testing problem and multiple comparison problem and the Bonferroni method, nor would we use that.
One second, I think my computer is frozen.
[00:33:00] [00:34:00] [00:35:00] [00:36:00] [00:37:00] Hello everybody.
My apologies.
My computer just crashed I had to restart and get everything [00:38:00] set up again.
I'm not sure what happened.
Is everybody still here? Sorry about that.
Well, thank you for your patience.
Again, that was not supposed to happen, but yes.
Back to the question at hand.
This question was again from Larissa Samson and the question was generally about the multiple testing problem.
I was still recording.
You can double-check that.
Yes, we're still recording.
[00:39:00] Screen-sharing here.
Actually the idea of multiple testing and multiple comparison problem is really exemplified in this one expression.
The idea is this, whenever we perform a hypothesis test or report a confidence interval, [00:40:00] we specify some level, Alpha.
Now, that level is used to 0.
05, probably for historical reasons more than anything.
But the idea is that this is the probability of a type 1 error.
Now, those of you who remember your intro stats, type 1 error is when we reject our hypothesis, when it's true.
The problem is that if we then perform more than one test or put more than one interval, well, the chances that we get at least one type 1 error, those are going to increase.
Something similar goes for confidence intervals.
Because when we report our 95 percent confidence interval, we're saying that under [00:41:00] the procedure that we're using, there's a five percent chance that our competence interval miss.
Right now in this case, the procedure is a whole bunch of steps starting with collecting the data and then computing the data summaries and reporting the interval, that's our procedure.

We know that that procedure incurs a five percent chance of a miss.
Now, that means that if we report more than one interval, the chance that at least one of them will miss will add up.
It's like we're playing a lottery.
We are buying a ticket, well, except if each ticket has a five percent chance of winning and so if you buy 20 tickets, you have about two-thirds chance of winning with at least one ticket.
Well, except in this case, winning is bad.
[00:42:00] This expression puts it more explicitly.
Here C_i is basically that's essentially the competence statement, saying that the Mu_1 is between two and three, then maybe Mu_2 is four and five.
If we say that each of these intervals has a, say, 95 percent confidence, which means that Alpha_i here equals 0.
05, well, then we can say that the probability that every interval we report is correct, that is, it fits over the true value, well, that's one minus the probability of at least one misses.
[00:43:00] Now here we can actually do a more complete calculation using a somewhat different method.
Maybe I should use a document camera for this.
You know what? Yes, I think let me derive this with a document camera.
I think this is a little bit outside the scope of the course, but I think it will be helpful in the future for your future work.
One second, I think now I can't see the document camera.
[00:44:00] One thing after another.
Let me try something else here.
I'm going to share a different window.
This will have to do.
We still have our probability.
[00:45:00] We have C, the hypothesis test to interval.
The probability that they're all true, and this is called the De Morgan's law.
[00:46:00] Probability that at least one is false.
Now, here's the thing.
Let's make an assumption that they fail or don't fail independently.
Apologies.
I was hoping to do this on paper, but this is a bit harder.
Well, let's do it like this.
The probability that if they're independent, then we can write this as the product of the probability [00:47:00] that they're all true, that they're truly multiplied.
That's the probability of the miss.
Now, if all the Alphas are equal, then we can say that that's one minus Alpha to them if all the Alphas are equal.
With the interpretation of this is, is that if you report a whole bunch of 95 per cent confidence intervals, [00:48:00] the probability that none of the misses is going to be 0.
95 to the power of the number of intervals in the report.
That can shrink pretty quickly.
I'm going to pull up R now to illustrate that.
Actually, I need to show different window here.
This is just using the plot function.
Maybe I should've used the RStudio [00:49:00] and I'm pulling up another version of R here.
This is what I want to show you.

If you report that is that if you only have one confidence interval with 95 per cent confidence, then we can be 95 per cent confident that it's true.
But if we have two of them, all the probability that both of them are correct is now only 0.
9 and so on it declines.
It doesn't decline linearly, it declines exponentially or like this.
By the time we have 20 confidence intervals, we are actually have more than an even chance that at least one of these intervals will miss, so we're no longer 95 per cent confident.
Now, what [00:50:00] methods like Bonferroni do, and I'm going to come back to the slides now.
I said I'm going to come back to the slides now.
Now this is another way of deriving this bond that is a bit more conservative.
The idea is that we have this probability of at least one of them is false.
That probability is going to be less than the sum of the probabilities that each of them is false.
We can actually see that in that graph that I showed you.
Because the idea is that if more than one of them is false, that's almost like wasted probability.
[00:51:00] You have this inequality and that means that we can write this as one minus the probability that they're true summed up.
This one minus probability it's, yeah, it's true that's the Alphas so the idea is that there is the miss probabilities.
The idea then is that if we want a confidence level of one minus Alpha, like just the sum experiment like 0.
05 for all of these combined, then what we can do is we can just pick Alpha i equals Alpha over m.
This is called the experiment-wise Alpha, Alpha over m.
Then that will guarantee that the probability that all of them are true, all the confidence intervals to cover is in fact going to be at least [00:52:00] one minus Alpha.
That's wonderful needs of adjustment.
Now the other approach and this will also motivates why we need this adjustment.
Because the errors accumulate.
Now, the confidence regions for based on these types of statements, based on the t-distribution, but mostly the f-distribution, the idea here is that once we compute this expression, we can plug in any number of l's here and all of the confidence intervals put together will have combined experiment-wise coverage of 1 minus Alpha.
The idea is that we might have a confidence interval from Mu 1 and then a confidence interval for Mu 2 and then a competence [00:53:00] interval for Mu 1 plus Mu 2.
Then a confidence interval for Mu 1 minus Mu 2 and any other linear combination, we report them all and are Alpha errors do not accumulate.
Now there's a price we pay for that, which is these intervals are typically pretty wide as a result.
It's a trade-off.
Now, when do we use Bonferroni? Generally, when you have relatively few confidence intervals, you report.
Because in practice, we often don't want every single possible linear combination of the elements of the mean vector.
Usually, we want something more modest like maybe the confidence interval for Mu 1 and a confidence interval for Mu 2.
Now that's only two of them, which means that instead of computing 95 per cent confidence interval, you compute a 97.
5 per cent confidence interval, that is [00:54:00] with Alpha 0.
025.
Then that'll give us a total confidence of 95 per cent.

I don't know how much it helped.
It is a little bit convoluted, but it does come back to this background of what do we do when we actually report a confidence interval.
I really hope the document camera will come back at some point.
That's that.
The next question is a series of questions from [inaudible 00:54:54], and the questions are, how does spectral decomposition, how we're going to use it in [00:55:00] this course? Also, compare or contrast diagonal matrix, symmetric, orthogonal, and real versus complex eigenvalues.
For part of that, it would be great if I could get the document camera working.
You're right, we're about the hour, so good idea.
I almost didn't notice, Alex, great idea.
Let's take a 10-minute break.
I will log out, so if I disappear, don't worry, I'm just going to try to restart the browser and get the camera plugged in and see if that works.
But great idea.
[00:56:00] [00:57:00] [00:58:00] [00:59:00] [01:00:00] [01:01:00] [01:02:00] [01:03:00] [01:04:00] [01:05:00] I'm back, so good to start by document camera.
[01:06:00] Let's go.
Good.
Welcome back, everybody I think it's been about 10 min.
The first question is, how does spectral decomposition applied to this course? The answer is, one answer is one week, but that's where the glib answer, but because it shows up in the principal component analysis.
But for now, I'm going to try to provide a general description.
The kinds of decompositions we're interested in are specifically of [01:07:00] symmetric, non-negative definite matrices.
Conveniently that's the property of variance-covariance matrices.
When you have a distribution and you will compute its variance-covariance matrix, that matrix is going to be non-negative definite or is the positive definite if the variables are linearly dependent, that is if you cannot construct any of them as the linear combination of the others.
None of the variables are degenerate in the sense of they only have one value.
But again, if the variables are reasonably well-behaved, then you have a positive definite matrix.
Then it turns out that eigendecomposition tells you quite a lot about the [01:08:00] structure of this matrix.
Again, for example the first eigenvector, that is the one that corresponds to the largest eigenvalue, that corresponds to the most biggest direction of variation in the data.
In fact, we began to prove it here.
In the next week, we'll be talking about principal component analysis, which involves solving something very similar to this.
In fact, it turns out that the first eigenvalue is the amount of variation explained by the first principal component.
This is why this is important.
In Week 6, when we talk about clustering, in particular model-based clustering, we will find [01:09:00] out that this decomposition of a covariance matrix can be used to specify the shapes of our clusters and their sizes and their orientations and all sorts of properties for those clusters in a very convenient parsimonious manner.
In the context of tests of covariance matrices, you want to know e.
g if the true covariance matrix for a number of variables is a diagonal matrix as opposed to some other

matrix.
The eigenvalues are often, the test statistics are in the form of the eigenvalues.
Similarly, when we do multivariate ANOVA tests and that include both canonical correlations and multivariate linear models.
I think that's Week [01:10:00] 3.
In Week 3 it turns out that a lot of the test statistics that we can use are going to be functions of eigenvalues of variance-covariance matrices.
Really this decomposition is going to show up everywhere.
I hope that answers that part of the question.
I'm going to switch to hopefully document camera.
Yes, good.
Document camera.
Position it so that I can actually use it.
[01:11:00] The next question was also by y k.
Y k is to talk about what compare contrast diagonal matrices, symmetric, orthogonal and real versus complex.
There is that line I want to use a document camera here is because it makes it easier to just draw these matrices and talk about them.
Diagonal matrix is pretty straightforward.
It's a matrix with elements a_1, a_2, etc, up to a_p that's showing a_p by p matrix [01:12:00] on the diagonal and the zero elsewhere.
Now, it has a lot of very simple properties.
For one thing, it's a determinant is always just the product of the diagonal elements.
If you multiply a vector by that matrix vector, you end up multiplying each element by the corresponding element of the matrix.
A_1 x_1 through a_p x_p.
Pretty straightforward.
Now, the asymmetric matrix has some useful properties as well.
In our case, we're mostly concerned with, [01:13:00] it has a diagonal and then curve, and then these parts are symmetric.
One nice property it has is that its symmetric matrix a is equal to its own transpose.
That can be convenient for various derivations.
But the property we were talking about that if the matrix is symmetric, all of its eigenvalues are real.
Again, this is a result I probably won't be able to prove off the top of my head.
But the nice thing about it is that if it means that when, well, if we're going to do a spectral decomposition of a variance-covariance matrix, which again are always symmetric, we'll only get real eigenvalues.
Speaking of which, again this comes back to diagonal matrices.
One thing it represents is that [01:14:00] if it's a variance-covariance matrix, then that means there's no correlation between the variables that are uncorrelated.
Now, does that mean they're independent? Any thoughts? I have disagreement.
Anybody wants to make an argument? I'll give you a hint.
I believe it's the software demonstration for multivariate normal distribution.
There is an example [01:15:00] of two random variables that are uncorrelated, but are in fact not independent.
The answer is not always.
Now if the variables are jointly multivariate normal, then yes, uncorrelated, this implies independence, but only then.

Well, there might be some other distribution families where it does, but it's only guaranteed for multivariate normal.
That's a bit of overview of diagonal matrices via.
An orthogonal matrix is not going to be symmetric typically, in fact, [01:16:00] it's hard for it to be symmetric because its transpose has to be its own inverse.
The idea of an orthogonal matrix is that its transpose is its inverse.
This is a definition we talked about last time a little bit.
Definition of the inverse is that the product of a matrix with its inverse is the identity matrix.
That's the definition of matrix inverse and it works in both directions for an orthogonal matrix.
[01:17:00] From that, it has certain properties and I think we discuss those properties before.
Those are the different kinds of matrices.
Is that what you're looking for YK? Great.
Next item.
This one is from Sarah Ray and the question was, can we derive the multivariate normal density? The answer is yes.
Now, here it gets a little bit of a dilemma because you see, the question is, how do you actually define the multivariate normal distribution? I decided to not get into [01:18:00] in this class.
Here's the thing.
Really, what is the multivariate normal distribution? On what grounds do we say that these variables are multivariate normal distributed and these are not? On what grounds? Well, it turns out that the definition that is used in formal statistical theoretical derivation is what's called the Cramer–Wold argument.
That argument basically says that a random vector X.
I'm going to use this for vector notation if you don't mind.
It's used in other places as well, but for the purposes of this, I'm going to use [01:19:00] this.
Actually, no, I'm going to use the underlying like that.
Random vector X, it's going to be normally distributed with some parameters in three dimensions.
If for all constant vector C does not equal to zero, C transpose X is distributed normally.
Now, C is a vector and X is a vector, so if we take C transpose X, we get a scalar.
We multiply [01:20:00] the corresponding elements of them and them and add them together.
If the result is normal, then for any possible C, we can write down that isn't just a vector of zeros, then, in fact, our X was jointly multivariate normal all along.
Again, this is called the Cramer–Wold argument.
I'm characterising a distribution by its linear combinations.
Now, what does that mean? Well, it means when we're thinking of it is that if you take two variables that are jointly multivariate normal and you add them together or subtract them from each other or multiply them by some constant numbers and then add the results up or any other linear or affine transformation, [01:21:00] then you will get a univariate normal distribution back.
That holds for more than two variables.
Now, turns out that from that particular definition, you can derive the variance of the multivariate or you can derive the variance.
You can derive the density of the multivariate normal distribution and you can derive a whole bunch of other properties of that distribution.
But again, the problem is that derivation requires things like characteristic functions and we're not getting into that in the time we have.
Instead, what I think I want to do, is I want to approach this from a different direction.
I'm almost the converse of this.
Let's say that we have [01:22:00] some vector Z and that vector is distributed.

Maybe that's not great way to write this actually.
We can write it like this, Z_1 through Z dot, dot, dot Z_p are independent identically distributed normal.
The vector Z is just all of these put together.
We have standard normal distributions all stacked in a vector.
Makes sense? We started with Z, so then you have Z now and so that's our first ingredient, so [01:23:00] our vector Z.
Our next ingredient is the fact that we have a density of the standard normal distribution.
This density is of course, 1 over square root of 2 Pi e to the minus z squared over 2.
Here this is normal with mean 0 and variance 1.
That's our second ingredient.
Our third ingredient is going to be two matrices.
One of them is Mu, or let me treats Mu as a vector and the second one is matrix Sigma.
[01:24:00] One thing to keep in mind is that remember, one thing we could do with eigenvectors and eigenvalues or with eigendecomposition is we could take a square root of a matrix.
Remember that? Well, we have our Sigma to the 1/2.
We're going to say that Sigma is symmetric positive definite, because otherwise you can't actually write down the density of the multivariate normal distribution.
These are the ingredients.
Now, let's consider the variable x, which is defined as Mu plus Sigma.
This is Sigma to the 1/2 and minus 1/2 just let me clear.
But we'll be using both of these.
[01:25:00] Sigma to the 1/2 times Z.
What can we say about this variable? Well, we can say that elements of x are linear combinations of elements of normal distributions, so they're probably going to be normal and similarly there are various other properties.
But also what we can say about that is Sigma must be non-singular, because like I said, otherwise you cannot actually write down a density for normal distribution.
Now turns out that you can still define a multivariate [01:26:00] normal distribution, even when Sigma is singular.
In fact, this original definition right there, the Cramer world argument that you can still use that definition there.
You might have to tweak it a little bit because maybe not all this produce normal distribution, but actually, no, I don't think you need to tweak it at all.
But this definition works even if the Sigma is singular.
But again, when if basically in order to derive it that way, I have to introduce a mathematical tools that we just don't have time for and also involve complex integration.
But yes, so that's our condition.
[01:27:00] Then the expected value of X, well, that's pretty straightforward.
We looked at some properties of moments I posted, although I think they might not actually be showing up correctly in Chrome browser.
Let me know if they don't.
We have the expected value of X.
That's going to be the expected value of Mu plus Sigma minus 1/2Z, which we can write as Mu plus Sigma to the 1/2 expected value of Z.
Now, each element of Z is just normal, with the expected value of 0 and variance 1, so this goes away.
Now we've established that the expected value of the resulting variable is Mu.
Now, what about this variance? I'll just derive [01:28:00] that.
That should be pretty straightforward as well.

Now, for the variance now the constant bit doesn't matter, so really the variance of X is going to be Sigma to the 1/2.
The variance of Z, Sigma to the 1/2 transpose.
Now here I'm going to cheat a little bit because in our definition of the matrix square root, now you might recall how we define it.
Let's remember that that's actually just was P Lambda to the, right this is [inaudible 01:28:52] and then the square root of 1/2, P transpose.
The idea is Lambda is a diagonal matrix, [01:29:00] so we can just take square root of the elements.
But the thing is that it's obvious that this is a symmetric matrix.
Really the transpose doesn't matter and this variance, that's just the identity matrix because it's independent with variances zeros, so really we end up with Sigma 1^1/2 times Sigma to the 1/2 and that's Sigma.
We have a distribution that again we can't prove it very rigorously, but should be about normal and has mean vector Mu and the variance vector Sigma.
The idea is that now we have a construction for this distribution.
Now this is a transformation, and this is where we bring in one more tool and that tool you will find in Week 0 multivariate probability revision [01:30:00] under transformation formulas.
The idea there is that if you can write one random variable as a function of another random variable, ideally of the same rank, then you can write down its density function using a very specific formula.
All right.
We have our function of that so let's begin by that formula.
By the way, on the slide.
This is formula 0.
19 or 019 rather.
It's written in terms of y and x, and I'm going to write it in terms of x and z.
[01:31:00] Let's say we have the sum function here and we'll specify that function momentarily.
Then the joint density of X and again, I'm using the equation 0.
19, but I'm translating into the notation [01:32:00] we're using here.
This density, of x_1 to x_ n or x_ p, rather.
X_p i going to equal to, well, what we're going to do is we're going to essentially solve for z as a function of x.
The inverse of that, we will plug it here and plug it into the inverse of this transformation and plug it into the density of z, that is the variable we're starting with.
F_ z of one z_1 of x_1.
Actually, I'm just going to use x vector just to save us writing.
[01:33:00] But that's not all.
There's also one component here, which is what's called the Jacobian.
I will again reproduce it momentarily.
Again, I'm going to use just more concise notation.
What is this Jacobian? Well, it's the determinant of the Jacobian matrix, which is the derivative of, well, we could write a number of ways, but the Jacobian of x here is going to be the determinant of the matrix of derivatives of z with respect to x.
That's [01:34:00] one way of writing it.
Now, because of what is called the inverse function theorem, the multivariate version of that.
We can also write that as derivative of x with respect to z inverse.
But fundamentally this is a matrix of derivatives so that actually I can show you this is not the term, this is the absolute value.
I misspoke.

Now, this expression is a little bit ambiguous, but the idea is that we're going to take the derivative of z with respect to x, [01:35:00] and then evaluate it at the corresponding values of x.
Now, how do we get the derivative with respect to x? Well, it turns out to be pretty straightforward because what we need to do here is we need to start with this expression and we need to solve for z.
Let's do that.
First, we subtract Mu from both sides.
Then we pre-multiply both sides by Sigma to the minus one-half.
That is, the inverse is similar to the one-half.
We get on the other side we have said, make sense? We just solve for z.
[01:36:00] Now we can do it, we can differentiate z with respect to x, and it's pretty simple because really we have z equals some constant times x and then minus some constant which goes away.
That implies that the derivative of z with respect to x.
These are all determinants of these matrices.
But here I'm just going to write down the derivative itself and that's just going to be sigma to the minus one-half.
Now we could have also started with this expression, computed the derivative of x with respect to z, and then inverted the result that would work too, or took the determinant and inverted that, that works too.
[01:37:00] But in any case, now we know what the Jacobian is.
We know that It's going to be the Sigma to the minus one-half determinant.
Now, here, one thing, if we do take the determinant of this, well, one thing we can do is we can see the determinant of the product to the eigenvalues, we can actually use the decomposition to show that we can actually write it like this.
That might look a little familiar if you.
[01:38:00] Okay, so now let's put it all together.
What is the joint distribution? We start with the joint distribution of all the Z's.
Now, they are independent and each of them has a density that looks like this.
There should be an i here.
So the joint density of all z's put together is going to be 1 over 2 pi to the P over 2.
Then here we have a product of this, e to the minus zi's.
So we'll just add them and the exponents.
So [01:39:00] e to the minus sum i equals 1 to p of zi squared over 2.
So we know there's a z, we know the Jacobian.
We know how to go from the vector.
We know how to go from x to z.
We know if given x we can plug in z.
So let's do that but first actually we are supposed to do one more little thing here.
So this is 2Pi to the P over 2.
Then here let's write this as minus one-half [01:40:00] and instead of the sum let's use use a dot product.
In other words, let's write it like this.
So now let's actually plug in.
Let's put altogether.
So first we have this expression, the constant expression 2Pi, P over 2, then we have this expression.
So let's actually begin substituting things in.
Now here we have z, so let's substitute in z here.
So we have sigma to the minus one-half [01:41:00] x minus Mu vectors transpose sigma to the minus one-half x minus mu.

So we just plug this expression into here.
Then we have the Jacobian.
The Jacobian we know is sigma to the minus one-half.
So this sigma to the minus one-half, let's put it down here.
So we have 2pi, p over 2 and then that's going to be the determinant of sigma to the one-half in the denominator.
Then you have e to the minus one-half and then here, [01:42:00] well, the transpose of a product is the product of the transposes with the order reverse.
So let's do that.
We have x minus Mu vectors transpose sigma to the minus one-half, transpose sigma to the minus one-half x minus mu.
Well, then remains to do one thing.
Remember these are symmetric.
So really these are just sigma inverse to the minus one.
That's the multivariate normal density function.
[01:43:00] So in some sense, the assumption we started from is that the multivariate normal distribution is basically an affine combination of independent standard normal variables.
There are lots of little elements involved in order to come together.
So there's a reason why all these bits are here.
[01:44:00] Any questions? So we've got about minutes minutes left.
So I guess I'm going to double check no keys.
I don't have anything else queued up to go over.
Is there any other questions about anything? I'm happy to answer them.
[01:45:00] [01:46:00] You're all thanks for sticking around to the long proof, now it's late.