[00:01:00] Everybody welcome, thank you for coming.
The plan today is, I will first talk a bit about Quiz 1 and provide a few more hints, and then I'll talk about the question that was asked on the thread about when do we use 1 minus Alpha and when we use 1 minus Alpha over 2.
Then I'm going to give some hints about check your understanding questions, you go through one or the other, check your understanding question solution in more detail, and I will, of course, post the solutions on Wednesday as always.
Then I was asked to go over one of the slides, so I'll do that and then we'll have a free form Q&A.
Any questions before we start? [00:02:00] No worries.
A few hints for the quiz, one question I got was actually from the last slide, this one.
This was a series of questions universal versus normal only.
Here's the thing, there's a reason why we use multivariate normal distribution.
We use it because if our data can be well approximated by the multivariate normal distribution, that means that [00:03:00] there are certain properties of it that we can use in our analysis.
But the same time, there are certain properties and laws and statistics that are always true regardless of distribution.
The goal of this exercise is to consider which of these properties are always true, so we can all rely on them for which ones do we need to check for multivariate normality? This will disclose to the specific question.
In this case, normal means that this property holds for normal distribution and it may hold for other distributions, but not always.
[00:04:00] On the other hand, always means that, yes, this always holds for any combination of variables provided a mean and a variance exist.
These are some properties that we've looked at.
Some properties and in various parts of properties of normal distribution and elsewhere in your notes, you will find answers to these questions.
That was one question.
Another question, and this was more of a logistical one and I might pull or send out an email about that, so you were asked to derive functions and somebody very astutely asked, do we also need to specify the range or the domain of the function? This is f [00:05:00] of x_2, so should we also specify the range of x_2 in this function and the range of x_1 in this conditional density function? The answer is no, you don't have to, and in fact, if you do, it might actually confuse.
There was also a quick question about this slide or this question, and in particular, sorry, not this question.
This question.
That's better.
[00:06:00] My apologies, I think I'm having some system issues, hopefully, it won't crash like the last time, let me just maybe close applications to see if I can free up some memory.
To be back in business.
Here, the question was in particular, one question was, so here, we have our confidence ellipsoid and the question is A, B, C, and D here, what do I mean here? What are these points supposed to be? The answer is, yes.
Well, the answer is yes, these are the end points of the major and the minor axis of this ellipsoid.
Now, as a matter of fact, I have shown you how to compute them [00:07:00] and this is probably going to be the biggest hint today, but hopefully, it will just give me [inaudible 00:07:27].
As you know, our studio is very unreliable, I've contacted the support several times, but these two haven't been able to fix it, I'm sorry about that, but there's nothing I can do about it at this point.
That's it.

[00:08:00] Here, notice that whatever code is up here has actually produced a confidence ellipsoid graphic, that with the major axis and the minor axis or major axis and minor axis.
For this ellipsoid, so whatever that code is doing, it seems to be doing the right thing.
I would recommend trying to understand what that code is doing and maybe working off that.
There was a distinction in the question between the case where the co-occurrence matrix is known and the words not known.
I think I gave a hint about it before, but again, the question is, [00:09:00] I guess I'll give you a hint that the critical value, that is the 1 minus Alpha multiplying might be different and different house for you to figure out.
The Pinterest went blank.
Well, that's all the hints I have for the quiz, does anybody have any questions? [00:10:00] Sorry everybody, how long was I out for? Good.
Yeah.
At [00:11:00] least a standard model, the standard halting room crashed, so sorry for that.
Well, any questions about the quiz? Hi Sarah? Hi Powell, can I ask a question on the microphone? Yes, go ahead.
Yeah, so the distributions will also vary, I'm talking about the ellipsoid question when Sigma is known and the Sigma is unknown, so the distributions will also vary, right? [00:12:00] Yeah, so that's what I meant.
Yeah, thank you, that's all.
Thanks.
Glad to be of help, any other questions about the quiz? I'll move on to the other bit.
The next item was a question from the rest of Simpson and it concerned the question taken more broadly is when do we use a 1 minus L for critical values and when do we use 1 minus Alpha over 2? For that I'm going to turn on the document camera.
[00:13:00] Fundamentally, we have lots of expressions that they generally involve taking one tile of the F distribution or some other distribution and the idea is that when you conduct a two-sample t-test or a one-sample or two-sample t-test, we use a t-distribution.
There's a t-distribution fact, if you have a proportion test, we would use a normal distribution, [00:14:00] but the point is we have some kind of a symmetric bell-shaped distribution and we have our z-score or t-score and you want a two-tailed test, we want it.
We look at the p-value, either greater than z or less than minus z, so this is a say, proportion test.
Similarly, if you want to have a confidence interval, you pick a value, we'll call it maybe z star, such that you have Alpha over 2 in here and Alpha over 2 in here.
I think everybody's seen this idea at various points in their intro stats class [00:15:00] and then the idea then is that when we actually want to look up z star on the table or using a computer program, well, we know we have to have Alpha 2 below minus z star and f over 2 above z star, so therefore, while in here we have to have 1 minus Alpha because that's the confidence interval or confidence level.
The number we actually need to look up for a z is going to be 1 minus Alpha plus Alpha over 2, so from here and then from here, we end up with 1 minus Alpha over 2.
Or if the other way of looking at it as we have type Alpha over 2 above it, [00:16:00] so we have to have 1 minus Alpha over 2 at our z star.
Now, that's for when you're working with a t distribution or a normal distribution.
If you're working with a chi-square distribution or an F distribution, those distributions, well, when we're looking at it, is that the chi-square distribution, it's actually just the square of the z distribution, so what happens is that when you square the z or the normal distribution, we get a shape that looks like this, it's a positive number and let's not worry about its centre and spread.
But now, if your z star now becomes essentially z squared star, but now if you [00:17:00] have the same

square of z squared as greater than some value, well it could be that the z is above this value or below minus that value.

If we say the probability that the Alpha is now the probability that the square of the said distribution is above the z-squared.

Again, the other way of writing it is that we see that probability [00:18:00] that z is less than equals Alpha or equivalently greater than mantras minus Alpha.

Then when we square this it becomes the probability that chi-squared is less than Alpha.

When you're dealing with a Swiss type of distribution that is symmetric and has both positive and negative values, you generally use 1 minus Alpha over 2, whereas if you're dealing with a distribution that essentially only has positive, that it's the square of this distribution, all your Alpha gets put here, so this gets mapped here, and [00:19:00] this gets mapped here.

Similarly with F-distribution, F-distribution is like one of these and so in some sense it's automatically two-tailed like this, you only need to do the over 2 thing.

That's why when we're dealing with something like t-star or t actually denotation, you didn't notice something like this.

Here you would typically have 1 minus Alpha over 2 whereas for an F distribution you would have 1 minus Alpha for these values, so that's same thing with z, p, [00:20:00] and chi-squared.

Yeah, so that's a general heuristic, does that make sense? Yeah.

This isn't particular also why for some things like T tests and Z tests, you can choose to make them one-tail.

You can choose to only look at the hypotheses where Mu is greater than Mu naught or only those where Mu is less than Mu naught.

Whereas for chi-square tests, you don't get to decide on the direction it's all there anyway.

Now T-square [00:21:00] test uses the F distribution.

The central you cannot choose to have.

There's no such thing as a one-tailed [inaudible 00:21:14] T squared test.

If you wanted to test for difference in a particular direction, you could do that, but actually then you could reduce it to a plane T-test.

That's all under multivariate problem.

This is a general clarification I know that for some of you this might be a bi-trite, but it is something very useful to keep in mind and not forget.

The next title was some things about [00:22:00] questions 1E and 2B on the check interesting exercises.

The solutions will go up on Wednesday as always.

But before that, if you want to take a crack at it.

Fingers crossed it doesn't crush.

Here's the statement of the check your understanding question.

You have trivariate normal distribution with a given mean and a given variance covariance matrix.

It asks a bunch of questions about to [00:23:00] determine various things about it.

It was specifically asked about part E.

Justify that these three variables are independent of X1 minus these.

There's a number of ways to do that.

One way to do that is to calculate it out.

Because what you can do is you can say, have a new variable, the new variable or new vector, call it X1 prime, X2, X3, and X4, and new vector.

Then write down a matrix that maps X1, X2, X3, and X4 onto, X1 prime, X2, X3, and X4.

Then [00:24:00] go through this for the arithmetic control and that matrix the correlations between X1 prime and the other three is in fact zero.

Since we're normal distribution independence and uncorrelated are equivalent.

We can conclude that in fact they're independent.
That's one way to do that.
Another way to do that is to perhaps notice something about this expression.
It has a certain shape, you have a row vector times the inverse of a matrix, and then times the other three variables.
Now, what we have we seen this before, even better, we have X1 minus some [00:25:00] a row vector inverse of a matrix.
Then the rest of the variables.
Where have we seen this? Anybody? Well, we've got another hour and a half, so you can wait.
[00:26:00] I'll give you a hint.
It's on that big properties of normal distribution page.
Again, the tasks is just to justify that these two variables are independent of this expression.
The value of this expression which is a function of X1 and then X2, X3, and X4.
Now, the question is this pattern, X1 minus sum row vector, sum inverse of sum matrix, [00:27:00] then a bunch of random vector.
Where have we seen this pattern before? Yes, I believe that's property four though I must confess I don't actually have it memorised the number, yes, it is property four.
Yes, it's property four, and that property concerned the conditional distribution of a normal random variable given other conditional distribution over subset of multivariate normal random variables given the others.
There we had a pattern like this.
We had something about perhaps subtracting.
We had some idea that maybe this well, okay, you [00:28:00] can notice for example that, this vector 101.
That's maybe you might be able to find it in this matrix.
Similarly, this matrix, you might be able to find it in here.
Then you might end up with something that fits that expression.
We're taking X1, we're subtracting something that maybe involves mapping X2, X3, and X4 onto X1 or predicting X1 using those and go from there.
There's a hint for that again, I'll post the solutions.
The next one is there is this question.
I think this was just double-check that [inaudible 00:28:44] Yes 2B.
This one also 2B small saucer other.
Again, this style have a trivariate random vector.
We want to find a vector such [00:29:00] that X2 and X2 minus this are independent.
Again, I will give you a hint if you can work through this part, you immediately know the answer to this.
The next item was, let me pull up that page.
There.
The question is, we have this mu, mu Y, mu X1, mu X2.
Two variables: [00:30:00] Y, X_1 and X_2 and then we have a partition and then we have a partition matrix.
Actually, I think we should have made it more explicit to have three variables here.
The question was a request to go through the solution.
Let me work through that and if necessary, I'll switch to document camera.
First is the best linear prediction of Y using X_1 and X_2.
Now, that actually gets us back to our old friend property 4 and it's corollary when you have a univariate variable we're predicting and a multivariate variable we're predicting from.
Again, you have the Sigma inverse x, x.

That's this rectangle.
This square part of the matrix.
[00:31:00] Can I resize this? Yes, I think I can make it bigger.
Then Sigma naught, which is the covariance between the xs and y.
1 minus 1.
We could also write this as 1 minus 1 transpose and then times this matrix inverse.
We just get a vector.
Once we do that arithmetic, it's minus 1 and 1 minus 2.
Then we would also need to figure out what the intercept would be and that's pretty straightforward.
That line must pass through [00:32:00] the mu Y and mu X.
Again, we see that for the property 4.
I should plugin mu X for X, you will get and I can try to pull up that slide, but I'm not sure if I will.
Let me see if I can try to pull up that slide, maybe no.
There we go.
In this expression, you plug in mu 2 for X2 and you get mu 1.
We can use that property and basically plug-in and then basically just solve plug-in Beta star transpose, which again, this product times mu X.
[00:33:00] These have to be equal.
In order to figure out how much we need to offset this product, we just subtracted mu 1 minus Beta star transpose mu X and after some arithmetic, we end up with 3.
That's our best linear predictor.
We could also go directly to this bit and simplify.
This is a more direct way to plug in the numbers into that expression and solve.
Questions so far? [00:34:00] The next one was asking; how do we get the multiple correlation coefficient? There was a number of ways to do this, but one that we can use here is the two.
You could just look up the formula right here.
In fact, we have our Beta star and plug it into this expression and evaluate.
That's the multiple correlation coefficient we get.
Incidentally, this is also the mitral call.
It's the square root of the multiple R-squared for regressing Y on X_1 and X_2.
Then lastly, [00:35:00] the mean squared error of the best linear predictor.
The quickest way to do that is to take advantage of this fact.
We take the original variation of variance of Y.
This is the amount of variance explained by X.
We just take the variance of Y and subtract off the proportion that's explained by X, which gives us 7.
There's a number of ways to do that.
For example, we could do it from first principles, which is to use this expression.
But but, of course, the two expressions are equivalent in this special case.
Does that help? I think [00:36:00] this one was asked to by Conner Bigs who is not here.
Oh, what's the recording? I think the last item on the thread was from Jay.
The request was to basically go through the multiple correlation slide.
Let's talk through it, and then we'll probably take a break and we'll have free form Q&A.
[00:37:00] This is as it has transpired, is a pretty straightforward corollary of property for the multivariate normal distribution.
Because we have all these variables that we're using to predict one last variable, and if you want to minimise the squared error, then the optimal predictor would be the expected value of Y given X and when we have multivariate normality, then the function is linear.
Again, this is mostly a rehearse of what was on that slide.

Now, there's a number of ways to characterise how well [00:38:00] the x is predicted y.
If you remember your linear regression, the classic way there is to use R squared and as you probably already noticed, these definitions in some sense converge.
The idea is that we're trying to use some linear combination of Xs to predict y.
Well, but we know that the solution is actually found on this slide, are on the slide this slide, that is the slope that is used to predict [00:39:00] y with x.
We know which one is predicting the best one and so we can define the multiple correlation coefficient as a correlation between the x with which has been multiplied by, this coefficient Beta.
That is the regression coefficient or the coefficient that is best used to predict y from x and cross between that and the original variable, so in other words, I guess in the language of linear regression, that would be the correlation between y and y hat.
One way we can compute [00:40:00] it is by actually plugging things in now we know the covariance matrix of X and we know the displacement matrix of X and if we just, plug in the numbers, then we actually end up with this expression, just bunch of cancellation and simplifications.
This ends up being the correlation coefficient.
Again, the way we would do it, and if you'd like, I will do the derivation.
But the way we would set it up is we would again have this setup, this vector and this vector and write it as a linear transformation and multiply up to matrices and show that after everything cancels, [00:41:00] the covariance between y and y hat is going to be the square of this expression.
This is how we might talk about estimating it and the estimate of this quantity is we just plug in the estimates of Sigma and R.
The rest is again a lot like what we've seen in linear regression, the variance we started out with and the variance that's leftover, and we can also write it like this we can see that because if we take this quantity and square it, we just end up with [00:42:00] X squared and multiply it by Sigma squared y, basically, plug-in this expression into here.
We just end up with this expression, the variance leftover.
In terms of the interpretation well, if the x's are completely useless in predicting y, at least linearly predicting it then the coefficients will be close to zero and R squared will just be close to zero as well.
There will be no correlation between Y and Y-hat because x is not useful in predicting y, whereas if it's 1, then the only way that can happen is, if all the variation in y is explained by X, which means that it's just a straight line without [00:43:00] any errors.
Does that help? That concludes the answers to questions that were asked in the thread plus a few others that were asked about the quiz, either privately or in other threads.
[00:44:00] In that case, what I will do is I will take a 10 minute break and then I'll come back and just sit here in case somebody wants to ask questions.
[00:45:00] [00:46:00] [00:47:00] [00:48:00] [00:49:00] [00:50:00] [00:51:00] [00:52:00] [00:53:00] [00:54:00] I'm back.
If you have questions, please either type something in the chat or raise your hand, and I will attend to you.
In the meantime, I will have the camera and the microphone off.
[00:55:00] Hi, Amy.
Hi, how are you? Thank you, and you? I'm doing not too bad.
Thank you.
[00:56:00] Thank you for your response to my post on Ed.
I appreciate it.
I think I've figured out question 3.
I totally changed tact.
Can I ask you a couple of questions about it.

Please, go ahead.
Sure.
I was having a look at the lecture slides and I used an example that you did.
It was like an odd demonstration somewhere.
Does that sound like a good place to? No, I think I talked about it earlier, didn't I? I came in about 30 min into the session.
[00:57:00] I talked about it the quiz at the start of the session.
Okay.
Yes.
In fact, that demo should give you what you need.
Yeah.
Cool.
This is a question that's related to that then.
The excerpt from the textbook that I put in my post to you, I get different answers to that using the method that you have provided in that demo compared to using that method.
I was just a little bit unsure as to why.
Did I stuff it up? Is that supposed to be the thing? First of all, let me make absolutely sure I'm looking at the right thing here.
Yes, your post.
[00:58:00] I'm sorry.
Actually, which text is this? This is Else's.
[inaudible 00:58:24] Yeah.
The one that we've been working on with other examples.
Just checking it.
I haven't actually in a while so I don't have it memorised to last forever.
[00:59:00] The answers that I got were different.
From memory, I think the actual stuff that I would have inputted into R for each situation was different.
It was really similar, but they weren't exactly the same.
I was just wondering why this is incorrect.
What's wrong with this? [01:00:00] One quick note is that I think I don't remember whether it was your post or somebody else's post I answered about that, which was that when lecture notes use Chi-squared or f subscript, some probability they're counting from the bottom.
Yes, I recognise that.
I checked my answers compared to what they were getting.
I'm really just trying to think of what might account for discrepancies.
I think it might have been the fact that, if I look at your code or my code, so we take the x bar for example and add or subtract.
Then [01:01:00] they've got one of the eigenvectors times the square root of one of the eigenvalues times the square root of this scale factor.
I assumed that C in the textbook notation that they're using would be the scale factor.
However, it's the C that they're talking about in the textbook is different too.
Your code works based on a later version of the textbook.
I was just a bit unsure as to why I wouldn't get the same answer with this.
I think the textbook [01:02:00] does talk about the F ellipsoids later, right? Yes.
This excerpt is from Chapter 4, but yeah, in Chapter five, they discussed the F and the Chi-squared.
Let me see this discussion.
The difference in this one is that it's dividing by n whereas the textbook doesn't.
Yeah.

[01:03:00] It's okay.
On a deep conceptual level, there's a relationship between these two examples, but there are several differences here.
One is that it looks like in the example 4.
2, the variances of the two variables are equal because you can see that the matrix on the diagonal is Sigma_11 and Sigma_11 again, not Sigma_22.
Yeah.
That's one difference.
Another issue is that this is finding the contours of just a multivariate normal distribution, but the covariance matrix in that distribution is again Sigma_11, Sigma_12, Sigma_12, Sigma_11.
In your case, if after [01:04:00] the covariance matrix of potentially Mu hat or the X bar if you prefer.
Yeah.
That's going to be different.
In particular, that's good.
What is the variance of the sample mean as a function variance of the sample? Sorry.
Did you say something you're expecting me to answer or is that a rhetorical question? Sorry, can you ask me again? What's the variance of the? If you have a sample.
Yeah.
Just random sample and it there's observation of some variance, what is the variance of their sample mean? I'm sorry, I don't know.
I guarantee you that you do.
[01:05:00] Maybe I'm not asking the question well.
If you just have to say a confidence interval or test based on the t distribution, Z-distribution, your Z-score is what? For a sample proportion, it's p minus p_0 divided by what? That's not a great example.
Let me go to the t-test is it's x-bar minus Mu_0 divided by what? I can't remember.
[inaudible 01:05:43] well, [01:06:00] it's a sample standard deviation divided by the square root of the sample size, right? Yeah.
Yeah.
Does that ring a bell? Yeah.
Yeah.
In general, if you have a random sample with variance S squared, the variance of the mean is going to be S squared over n, and then square root it to make it a standard deviation.
That's the crucial difference in this case as well.
Because here this is just an old distribution whereas what you're dealing with confidence regions, is that you are dealing with the distribution of the mean.
I see.
Another thing, once again, I'm sorry if you've already spoken about this.
[01:07:00] I'm doing the final question, 7, 8, 9, 10, where you've just got to pick for all distributions or only for a normal distribution.
Where am I best to find the information to be able to help me with this.
I'm looking at the properties of the multivariate normal in last week now I should say.
This slide talks about multivariate normal however, in some instances, for some of those questions, I'm not sure if that applies to other distributions.
[01:08:00] I see what you mean.
That's a fair point.
Now, in some case, that page and others do talk about the properties in general.
This is what it means for the multivariate normal distribution.
The answers are in the notes, so some of them are not 100 percent given to you corrected.

Yeah.
I should essentially just keep looking.
Yeah.
I think that's everything then.
Good luck.
Thank you.
[01:09:00] Oh, sorry.
Can I ask one more thing? No, no need to apologise.
That's why I'm here, so go ahead.
With question 5, I can't remember what it is, hang on.
Let me pull it up.
That's regardless of the answer to the previous question [inaudible 01:09:41] Yeah.
For the very last part where it says conclude and you've got to choose multiple choice, can you just give me a bit of guidance here as [01:10:00] to the wording of, I'm trying to choose between two options and I'm just not sure based on the wording which is appropriate.
Yeah.
In some sense, this is not in multivariate analysis specific, [inaudible 01:10:27] Yeah.
The idea is you need to think about what the null hypothesis is, and what the alternative hypothesis is and whatever your p-value is, figure out whether you reject the null hypothesis or not and fundamentally the hint there is that we never prove the null hypothesis.
[01:11:00] We can retain it or accept it tentatively, but we never prove it.
We can only either accept it or fail to reject it.
Cool.
We accept or fail to reject? Yeah.
We only conclude or we only prove the alternate.
Yeah.
We only prove that it isn't something.
Yeah.
Great.
Thank you.
I think that clears that up.
Great.
Thanks so much for your help.
No worries.
Good luck See you.
Bye.