[00:01:00] We're now recording.
Quiz 1, what I'm going to do is I'm going to basically pull up a preview of the quiz the way I would see it and then we'll work through it, rather how you would see it.
Is the screen sharing correctly? It looks like it's sharing reasonably well.
The first question was a question [00:02:00] about probability distributions, particularly multivariate ones.
The idea was very basic.
There was a bit of an issue that I will show.
I didn't specify exactly how you were supposed to exponentiate things, and the preferred way to do that in Moodle is actually two stars.
If you lost marks because of that, please let me know.
I posted an announcement about that.
Everybody would have a slightly different problem here.
You would have a slightly different solution and I think you would actually get to see your own solution once you look at the review.
But, let's talk about the basic idea.
We have a joint density and we want to get [00:03:00] the marginal density.
That is, we want to know what is the density of just x_2.
Now, we`ll do that by integrating x_1.
Switching between document camera and screen takes a bit of time.
What I want to do is I'm going to go through the questions and then switch to the document camera and go from there.
We would set this up right now.
This dominant is you would integrate out x_1.
Here, what a conditional density.
The idea then is we take the joint one here, and we divide it by the marginal one so it's build up on part a.
Then the best approximation [00:04:00] one, I think I mentioned this a few times, the idea is that the best approximation that fulfils these criteria, that minimises the squared error is the expected value of x_1 given x_2.
Here again, you would need to set up the appropriate integral and evaluate.
Then lastly, given that, you would also want to set up another integral, and I'll write that in codon momentarily.
What I will do is I'll go through these two and then I'll walk through how I would have done them.
This one calls for a document camera, this one calls for r, what I will do is I'll switch the document camera now.
[00:05:00] The idea is that for part a, I'm not going to actually solve the integral because again everybody's answer is going to be a bit different.
But in fact, I would just use a computer algebra system because I'm lazy.
Then we have f, X_2 [00:06:00] of the x_2.
Can you see my writing by the way? Just want to make sure.
That's going to be always the integral.
In this case, let's write down the integrant first, f over x_2 with respect to x_1.
If we were dealing with discrete variables, we would be saying that for each of the x_2s that sum up over all the x_1s and to get the total probability of that particular x_2.
Then we have to figure out the domain here.
We're told that x_1 and x_2 are both positive, so zero to infinity.
Again, your answer might be different from mine because [00:07:00] your question was a little different

from mine.
If you review your submission on Moodle, you will see the final answer.
Right now this is for the benefit of those, I want to talk about how we solve the question.
Does that make sense? Again, your answer may differ from mine.
Well, here I am going to write out the first step.
But basically here, once you've computed this, you would say that f of X_1 given X_2 is just f of [00:08:00] x_1 and x_2 and divide by f of x_2.
Part c, and again, I'm going to just set up the integral, it's going to be the expected value of X_1 given X_2, which is in turn an integral from zero to infinity of x_1, f of x_1 given x_2.
Then finally d, we will compute, that is take the expected value of X_1 squared given X_2 equals [00:09:00] x_2.
Then we would just set up an integral like this one, but for the square, and then subtract part c squared.
That's a common calculation formula for that.
That's how you would solve that.
Any questions? For Question 2, maybe Sarah has a point.
I think I can post solutions or at least the code I would use to solve these in R, which is what I would use here.
Does anybody have any preferences? [00:10:00] Go over them here, I can post the code.
For these, in fact, it would be a bit of a hassle and I might as well posted.
We can do that.
But, I do want to talk about some of these conceptually because it's good worth discussing and may be yielding some questions.
The answer for these, you would just use the multivariate normality tests, either the ones described in lecture or others.
Sorry, go through week E materials or week [00:11:00] 2 materials.
I'll do that.
But, there are certain things I want to point out about the quiz before that and then we'll go through week 2 quiz.
Again, I think maybe I'm just going to post the code for answering this.
But, one thing to think about is the conclusion.
The correct answer is in fact b.
The p-value you'd get is pretty high, something like 0.
5 plus or minus, regardless of whether you do a pooled test or down pooled test.
The reason why it's B is, first of all, we don't reject the null hypothesis of equality, which means that we would say that there is not sufficient evidence.
Not my day to day, is it? [00:12:00] The polycentric mark question five, and there's a conceptual component here.
The p-value here would end up being somewhat 0.
52.
There's not sufficient evidence to conclude that the population is different.
This is different from concluding that sufficient evidence so they do differ, including there's sufficient evidence that they do not differ.
Because we don't ever really prove that two things are equal.
We can only prove that they are unequal, which means that I hear the null hypothesis is that they're equal.
The alternative is that they're not.
Now we would say that would you actually use B here rather than one of the others.
[00:13:00] Now for multiple choice once again, this is a better to describe verbally.

Now, true or false the joint distribution of two normal random variables is always multivariate normal.
The answer it's false.
Again, there was a demo about that where we actually had two variables that are normal but not jointly.
I think most people got it.
Now, I think these universal versus normal, I think a lot of folks have trouble with that.
Actually, in the final mark, I down-weight these because I'm not sure whether a lot of folks understood this question.
But okay.
But let's talk about this because some of these properties [00:14:00] of normal distribution are worth discussing and how they apply to these.
You have this expression, you break up X into X_1 and X_2.
This should really remind you that we maybe doing property four of the multivariate normal distribution.
What a function that predicts X_1 as well as possible.
That means to these quest prediction.
When we're told that this is the expected value of X_1 given X_2 which actually we just do that in question 1.
That's always true for any distribution released that has an expected value.
It is not unique to the normal distribution.
What multivariate normality gives you in this case is that this expectation expression would be linear in X_2.
[00:15:00] This is always true.
For the sample the maximum likelihood estimator of the Sigma hat is always what we're trying that is the MLE of Sigma, the correlation is basically the Sigma ij from Sigma and Sigma jj.
This one is also always true because if this is the MLE, then some function the MLE will be just that function.
This is how we translate covariance to correlation.
This is always true.
Again, this one was a bit challenging, so I've down-weighted in the final mark.
This expression now, this is a set of identities I believe will be posted [00:16:00] to the notes.
The variance of AX, A times X for some matrix X is just a Sigma A transpose, and it's always true.
Last if X_i and X_i are independent, the correlation is zero.
This one is actually true.
Now the opposite, which is that if Sigma ij equals 0, does that mean X_i and X_i are independent then turns out to be the answer is no.
Again, we have a software demonstration to show that.
But this one is always true.
I think that's it for quiz 1.
I think you're right.
I'll go ahead and just post the code for quiz 2 probably make everybody's lives easier.
Now let's talk about quiz 2.
[00:17:00] Let me share that window.
I think I can talk about there's only one question to talk about for the first one.
But just in case so I know that several have asked whether the quiz part will be weighted the same as the current.
The answer is, of course no.
True or false suppose that for random vector this correlation between [00:18:00] row 1 3 0, and then

the partial correlation equals to the ordinary correlation.
For this question, my hint would be, look at the formula for those in the notes.
You should have no trouble answering that.
The next part is, of course, the coding challenge.
Now this might be a bit unfamiliar to you, so if you have any issues, please let me know.
What I'm giving you is a variance covariance matrix, and we don't really care about the mean in this situation.
Now, I'm giving you a line of code that you can use to construct this matrix.
From that, I want you to use whatever tools you consider appropriate, you can implement the formula yourself or you can use one of the functions introduced.
[00:19:00] Or if you can find another R package that does it feel free and interesting to learn for me.
Yeah, once I go through the bits that I planned, I will.
Here, just use whatever tools are appropriate.
Again, of regarding quiz questions, I tried to reply to questions about clarifying the meaning of the quiz questions.
Yes, general help I prefer to use the webinar unless it's something really quick.
[00:20:00] The idea is to practice doing principal component analysis from the matrix.
You have the functions to use for the data that the can do it from data and this is an approach that uses the matrix all their functions from matrix as well.
Then just over here, I would recommend looking at the code in the demos on the code challenge and then you will be I'll uncovered that code night.
It should probably pretty much tell you how to [00:21:00] do this.
Now, one question was, what is the scaled versus unscale principle for analysis? The main difference is that in one case you're dealing with a covariance matrix, Sigma.
In the other, you're dealing with the correlation matrix.
Here, again just the way you would work with this as you would enter whatever your solutions are.
If I run a pre-check on this, it would tell me that it's happy with the numbers I entered here for these because it's a scalar which is five, because our forwards at all the others.
But it's unhappy about the vectors.
Once it's satisfied, then at least [00:22:00] you know that your answers in the right format.
Larissa, that is correct.
Yes.
I think both are used.
They're synonymous.
Vicky, you're right, it is confusing.
If you look at, for example, our output for regression and we'll talk about multiple R-squared, [00:23:00] in that context, people call it multiple correlation.
But total correlations is another way.
It's interesting, total correlation, maybe there's another meaning but maybe not, let me see.
It's a good question.
I see.
I think that there are different usages here.
I [00:24:00] think this is a good point, I think maybe what I should do is I'll probably update the notes to use one of the others.
Semi partial correlation.
What do we mean? Truth be told, I don't think I've used it before.
[00:25:00] The answer is no, it's not.
Now, I see why I haven't heard of it before.

It seems to be used in machine learning.
It seems to imply most of causal inference type of interpretation.
How about this? I'll look into it and get back to you.
I haven't used it myself, so I'm not very familiar with the idea.
Thanks for raising it.
But it's not on the planned curriculum anyway.
[00:26:00] That was an excellent question.
You can do either.
You can either just type your answers in.
Now, on the back-end, most of the popular packages for this course are installed, so you should be able to use them as well.
[00:27:00] Sorry, I think it's not switching.
There we go.
Go ahead, Sarah.
Good afternoon [inaudible 00:27:26].
Can you hear me? Yeah.
In the first slide of Week 2 where you were defining the partial correlation.
Here's the thing.
This was a bit sidetracked.
What I prefer to do is I want to go through the ones that were posted to the forum first, then additional questions.
No worries.
I'll wait [00:28:00] for you.
Sure.
I know we got a bit sidetracked because first two are questions about specific to the quiz and then the folks start adding questions.
I think chronologically, the first was actually a question about partial correlation.
The first question was about the partial correlations and was from [inaudible 00:28:43].
Actually maybe I'll pull it up here, one second.
I think this should work.
[00:29:00] Did that work? Good.
The idea of a partial correlation and conditional correlation is something we haven't really covered.
It's something I can probably explain later on.
It's not been a good day.
I think I'll leave this one for later.
[00:30:00] This, I think, applies to all of these questions.
The correlation is just something we measure on either a set of random variables or on a dataset.
In the context of a multivariate normal distribution, we can get additional information, additional weight.
Just as for multivariate normal distribution, the predicted value of one set of variables given another set of variables is a function of the covariances between those variables, or correlations between them and the variances.
Whereas for others, it's not.
That said, correlation is still useful for cases that are multivariate, normal, and in particular for multiple correlation coefficient, one interpretation that it does have that I think is a very [00:31:00] valuable one is that it is the correlation between the predicted value, is the Y hat and the Y.
Now, the relationship between the predictors, which are based on what you compute the y hat, that can be a linear predictor, in which case, it's optimal for multivariate normal.

What could be some other prediction? Or it could be not some non-linear effect of X.

Then in that context, R-squared or this correlation is in fact so meaningful.

It's still telling us how well we're predicting the outcomes.

But it's no longer linear expression.

It no longer applies in the multivariate normal setting and it no longer necessarily [00:32:00] has this simple formula.

The next question from that was the proof of this calculation formula for Sigma YY.

There are several things that we're talking about.

There's partial correlation, there's conditional correlation, those are two different measures of association.

In fact, several [00:33:00] different kinds of correlation.

Here, we're talking about whether they correspond to each other.

Actually, the general conditional correlation formula, can you paste the link to the Wikipedia article there and I'll come back to it later? I think maybe it'd be better if I could because I didn't get a chance to follow up on that.

Well, I'll keep it in mind.

[00:34:00] It's a thing.

I see.

To love total variation is something I don't think we've covered.

How about this? I'll probably use a document camera [00:35:00] later and, when I use a document camera for Part 2, I will also come back to that and explain what the correspondence there is.

I apologise I'm a bit less prepared than I would have liked to be for today.

But in the meantime, next question was, do all these tests require you that the random variables are normally distributed? The answer is no, if your sample size is sufficient.

This is generally true about most of the things we deal with in this course.

There are some exceptions, but for most things, it's [00:36:00] generally fine.

We know that as the sample size gets bigger, the distribution of the mean becomes more and more normal like.

But it's not just to mean, it's also the mean of the squared values in the data that also becomes more normal like.

It's not just that, it's also the functions of the mean or the continuous functions of the mean they become more normal like.

Which means that eventually basically lots of things that are not becoming normal like.

That means that thanks to the central limit theorem and something called the Delta method, and a few others pack results, if our sample size is big enough, they hold for [00:37:00] all distributions that have a mean and a variance.

Normal distributions have variance, for example Cauchy distribution, doesn't have a variance.

But for those that do, yes.

Now, how big a sample size you need? That really depends on the situation.

Some methods are more sensitive than others, for example, something like the t-test is pretty robust even for really small sample sizes.

But as we will learn later in next week or the week after when we talk about tests on covariance matrix.

Those tests can be very sensitive for normality, so there you yield a very big sample size.

Partial credit coefficients, you can test.

It looks like there's a typo, [00:38:00] this should be our row ij conditional and r_ij conditional.

I think the rest is correct, but now that I'm not longer sure, I'll need to double-check that.

Thank you for pointing it out.

Again I haven't had a chance to go back and look at that.

This bit, I'll take a look at it now come back to them for an answer it either on the form on Friday.
Then this question from Benjamin Costello, so Lemma 2.
1.
I'm not [00:39:00] having to do right now.
This is still Lemma 2.
1.
What's the point of defining it in terms of the second form? The answer is that Beta is Beta transpose, also Beta is supposed to be a column vector.
This expression will give us a row vector.
That's all there is to it.
[00:40:00] One more question about that Pavel, Yes, but let me first go through the rest and then we'll go through this question and then we'll come back.
Okay.
Finish these up.
This one is what's the difference between pcor c1,2R and parcor R1,2.
Parcor, get your matrix of partial correlations of each pair of variables given the rest of them.
Whereas pcor asks for specific correlation given other variables [00:41:00] which you also specify.
In this case, pcor c1,2R, will give you just a plain old correlation between the first and second variable.
Whereas parcor R will give you the correlation between the first and the second variable, given the rest of them.
If you wanted to do that with pcor, you would do pcor 1, 2, 3, 4.
I hope that helps.
Maybe I'll pull up RStudio and demonstrate.
This one looks like it just got posted.
Yes, this one is always conditioned on everything.
This one's conditional things you specifically asked for.
Let's take a five-minute break and [00:42:00] then we'll come back and work through some of these in more detail.
[00:43:00] [00:44:00] This is a very nice document.
Thanks.
Maybe I'll probably incorporate a link to it to future instances, of course.
Yeah, okay, but to the general point about the LMMSE framework, I guess the answer is yes.
So the multivariate normality, its relationship to this type [00:45:00] of framework is that these types of estimators are in some sense optimal under multivariate normality, and in particular also have additional interpretation properties.
Whereas for other relationships between variables, correlation indicates the type of relationship between two variables.
For multivariate normality that's actually a parameter in the distribution.
Does that make sense? [00:46:00] There could be other distributions for which this framework is also optimal.
Now, optimality is a more complicated idea.
But briefly, there are, as we saw with the idea of predicting one variable given the others.
For some pairs of variables, the relationship, the optimal prediction form is linear for others it's not we're going to see of others where it's not.
For normal it's linear.
It may also be linear for others.
But it is linear for normal.
We will encounter that in a few other cases.

If you have places where [00:47:00] we have okay, well this works more generally, but it is sometimes optimal for normal distribution.

It's the best you can do if your data are normal.

It is interesting that this was all for the core MIT OpenCourseWare.

No wonder it's so nice.

Anyway, so I think [00:48:00] now what I want to do is I want to switch to the document camera and I wish I had two screens here.

In lecture I would have two screens.

Again, this proof is not examinable, but yeah, you can actually get this result, which is pretty interesting.

But the idea is [00:49:00] plot matrix determinants.

That's not how I would derive it.

But yeah, it's interesting because you get this.

Basically we use this property of the inverse, and you can actually see that if you take the inverse of the correlation matrix and take its first element, then it's going to be there, the correlation matrix divided by the [00:50:00] determinant of the correlation matrix without that variable.

Anyway, I don't know, is it something that anybody is interested in me proving in detail? Yeah.

That's correct.

I don't know whether anybody is interested in me actually go through this in any more detail, but okay.

Well, I think maybe what I will do is then, okay.

Let's try it.

I'm thinking, what else do we need to use pencil and paper for and what else do [00:51:00] you want to go through? I think just okay.

I'm going to switch to the document camera now.

Let me go through that.

I can drop that.

It's weird.

For some reason, it always takes two attempts to get this thing to turn on.

I think one thing was this idea of the parallel between the conditional correlation and the partial correlation.

Conditional correlation or conditional covariance comes from where we use the property called the law of total variance.

I'm curious how many of you have seen this.

[00:52:00] But basically, if you have two variables, X and Y, we can write down the expected value of Y as or three.

First, there's law between the expectation, I think is actually something we want to talk about first here.

We can actually write that as the expected value of over the possible values of X, of the expected value of Y given X with respect to Y.

The idea is that, if you have two variables, X and Y, what you can do is you can essentially compute the expected value of Y given X.

Then you can take that expression for the expected value of Y given X, which is going to be a function of X, and take the expected value over the possible values [00:53:00] of X.

That will actually give you the original expected value of Y.

Now, it turns out there is a similar property for variances and covariances.

Here's what it looks like.

The variance of Y equals to the expected value over X of the variance of Y over the possible values given X plus the variance over X of the expected value with respect to Y of Y given X.

This is pretty interesting in a number of ways.

Again, it's a little bit outside the scope of the course, but [00:54:00] it is worth talking about I think.

The idea is that, the total variation in Y is the part of the variation that is explained by X.

This is part of the variation that is explained by X.

It is this one, plus the variation not explained by X.

What do I mean by that? Well, the conditional variance of something is going to be essentially, well, if you're thinking, in fact, let me draw a picture.

Let's talk about linear regression.

If these are bi-variate normal, then in fact, that's the relationship between Y and X.

We can think of our points, and we can think about having just the variance of the points.

Or we can think about the residuals [00:55:00] from the regression.

You'd have a point here and we can think about it's residual or we can think about it's essentially residual with respect to just the sample mean.

Now, if we have a line that fits well, then the variation around the horizontal line, it's going to be much bigger than the variation around the actual least-squares regression line.

It turns out that we can write this total variation around the horizontal line as a function of the variation around the least squares line.

That's this one, variance of Y given X.

Given that we subtracted over out whatever was explained by X [00:56:00] plus the variation over the X of the expected value of Y given X.

That's the variability essentially of this line, as opposed to a rounded.

Now, if that rings a bell with respect to the analysis of variance, that's because there is a very close relationship there.

This is the sum of squares total.

This is analogous to sum of squared error.

This is the sum of squared progression.

But of course, this looks more complicated than you usually deal with ANOVA, because under ANOVA, we assume equal variances.

For every X, Y has the same variance.

But [00:57:00] that doesn't have to be the case, which is why we have these extra bits here.

Imagine if in particular the variance of Y given X is constant for X.

Well, we can just ignore this expected value.

But if it's not, we can't.

For more complex predictions scenarios, you can't.

Now, how does that relate to these ideas of conditional correlation versus partial correlation? So first of all, there's a version of this that works for correlations, as you might expect.

You just replace variances with correlations.

But here you have the component that is explained by X, and you have what's leftover, the conditional one.

This you can view [00:58:00] as conditional variance, or at least it's the average conditional variance for this scenario.

In fact, for something like a normal distribution where this variance is just constant for all values of X, the expected value is not constant but the variance is, then the partial correlation will in fact be the same as the one that we would get from taking I guess the conditional covariance and dividing it by the variance.

Does that make sense? Again, this works for any distributions, assuming that they have expected values and variances, which is [00:59:00] not true for all of them, but it's true for most of them.

Does that make sense? Then the next item was, I think since we're already here, let's work through the whole correlation thing.

The whole thing comes from the result of this property that again X inverse, in this case we're just going

to say [01:00:00] just because it's just a special case we're interested in.
Just a general guideline.
If you're going to type the question, please click "Raise your hand" so that I would know that you're typing something because I might have moved on by then.
No harm done.
But yes, so here, the conditional correlation, one way to think about is just the conditional correlation here I think is conditional variance.
If these are vectors [01:01:00] then these are going to be matrices.
It still works perfectly well.
Then basically you could take this expression.
This would be a perfectly valid covariance matrix.
Then you could just convert to correlation matrix by dividing the rows by the square root of the diagonal, and the columns by the square root diagonal.
Yeah, that works.
Actually, no, I think maybe I need to fix the notation in this case because, let me just make sure.
Apologies.
[01:02:00] I should probably fix the notation because it's a little bit ambiguous.
Yeah, Sorry.
For some reason I'm not giving it the website is not quite working for me at the moment, but let me see if I can.
Yeah.
[01:03:00] There was a typo in the notes.
I'm sorry.
I thought we had fixed it, but apparently we haven't.
The way it should have been written is so here X11, which is a matrix.
I'm sorry, I just discovered there was an error in the notes that I should probably needs it to be fixed.
But okay, so the general property works like this.
[01:04:00] Okay, here, x i j.
There was actually an error in the notes.
I'm sorry, that was that.
Here's what I think happened.
There are two similar courses that we're teaching, and I think we fixed it in one set of notes but not the other.
I'm sorry about that.
Yeah.
Thank you.
The thing is that if in fact we go back to the expression.
[01:05:00] I guess I can go through that now or we can do that.
Anyway, would be basically using the manipulations that are given on the slide.
We end up with the expression 1 minus R equals sigma over our [01:06:00] y that's to those many.
Yeah.
But the thing is that the C is the covariance matrix between the Xs.
It just happens to be sigma with the first row, and the first column removed.
In fact, we end up blending onto this formula, this C, we can actually write that as sigma 1 1, or sigma 1 1 removed.
Then we can just go from there, because if you have the inverse covariance matrix, [01:07:00] let's see but then yeah.
I think it's pretty straightforward from there.

Once you can reduce something like that and observe that this just happens to be, I'm trying to think.
Here so I'm going to write this [01:08:00] as Sigma inverse 11 divided by Sigma squared YY.
Sorry, I just want to make sure it is correct.
Anyway, from that point on you can [inaudible 01:08:43] then think it's clear.
All right.
Moving on, let's see.
Sorry, it's not going well today.
[01:09:00] All right.
I think those are the ones that code for pencil and paper of the.
I think that's all that's on the forum except there was a question that will just post it by Sarah.
That no matter how strong correlation card help us identify direction of influence.
[01:10:00] This is a good question Sarah, the language is ambiguous.
The direction is talking about here is not the direction positive versus negative.
It's a direction of causality.
That's why I think the example we had was, what is the height, age? This was our partial correlation example.
Partial correlation example, the example was that, excuse me, my browser's frozen.
All right.
Intelligence weight and age.
[01:11:00] Here, we could do this coordinate intelligence test.
For example, we would not be able to tell from correlations alone whether there's a relationship between say, intelligence score, test score and age.
But we can estimate that it's positive from the data.
But we cannot fundamentally from the numbers alone, know that.
But we could infer that from the fact that somebody's age is something pretty exogenous.
So the causal arrow probably goes from age to intelligence and we made them into a story for how age affects the score of an intelligence test.
But that is not something that from the council statistical theory, but rather, I guess sociological or biological theory.
[01:12:00] But the direction of positive versus negative correlation, yeah that's just arithmetic.
I hope that makes sense.
Very interesting stuff from the MIT OpenCourseWare.
I think these are all the questions on the forum.
[01:13:00] I think there was another question from Benjamin I think.
Are there any other questions? Jay, we've gone through the multiple choice ones and we briefly went through the question one.
But the rest, I think what I'll do is I'll just post the R code I used and work from there.
[01:14:00] All right, any other questions? [01:15:00] If there are no more questions, what I will do is I will just switch into the consultation mode and just sit here and wait in case anybody has any questions.