Correlation

Testing relationships between two continuous variables

Correlation

Today's goal:

Teach you about correlation, a measure of relationship between two continuous variables

Outline:

- Explain covariance and correlation
- Demonstrate how to compute correlation in R
- Discuss advanced forms of correlation (which lead in to our next topic, linear regression)

Covariance and correlation

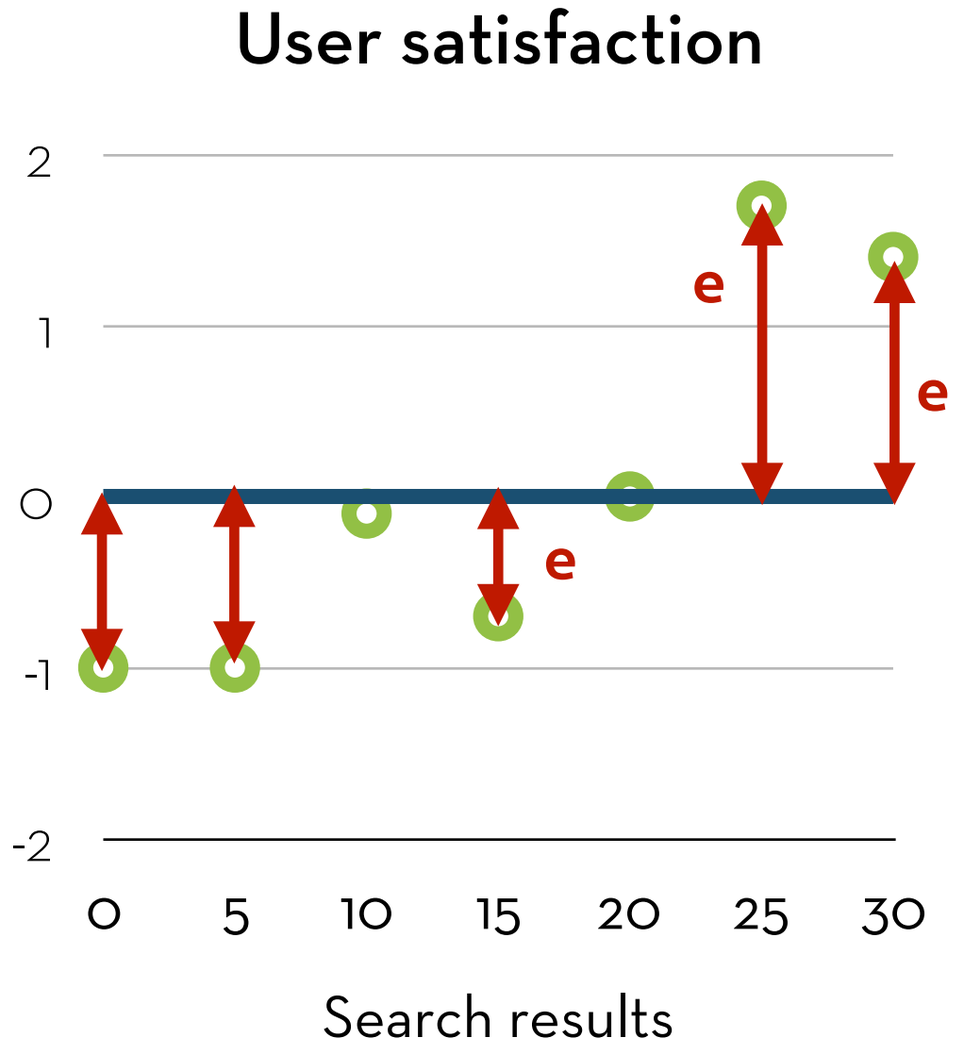
An explanation

A quick recap

Variance is the variation of the data around a model (e.g. the mean)

$$s^2 = \sum(x_i - \text{mean}_x)^2/(N-1)$$

It is the sum of the **error in x times the error in x**, divided by the degrees of freedom



Covariance

Covariance measures the relationship between the variations of two variables, x and y

$$cov(x,y) = \sum(x_i - mean_x)(y_i - mean_y)/(N-1)$$

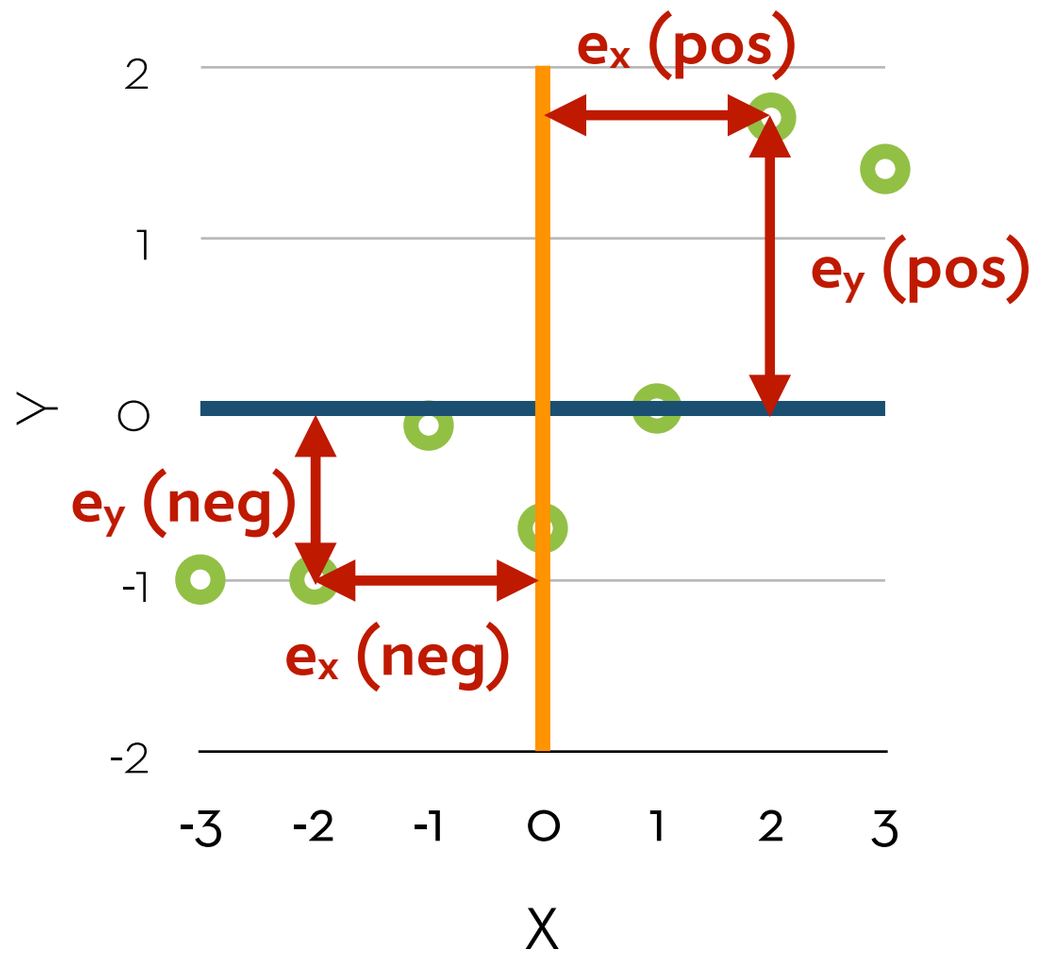
It is the sum of the **error in x times the error in y**, divided by the degrees of freedom

Covariance

Covariance is positive when:

The error in x and in y are both positive (1*1 = 1)

The error in x and in y are both negative (-1*-1 = 1)

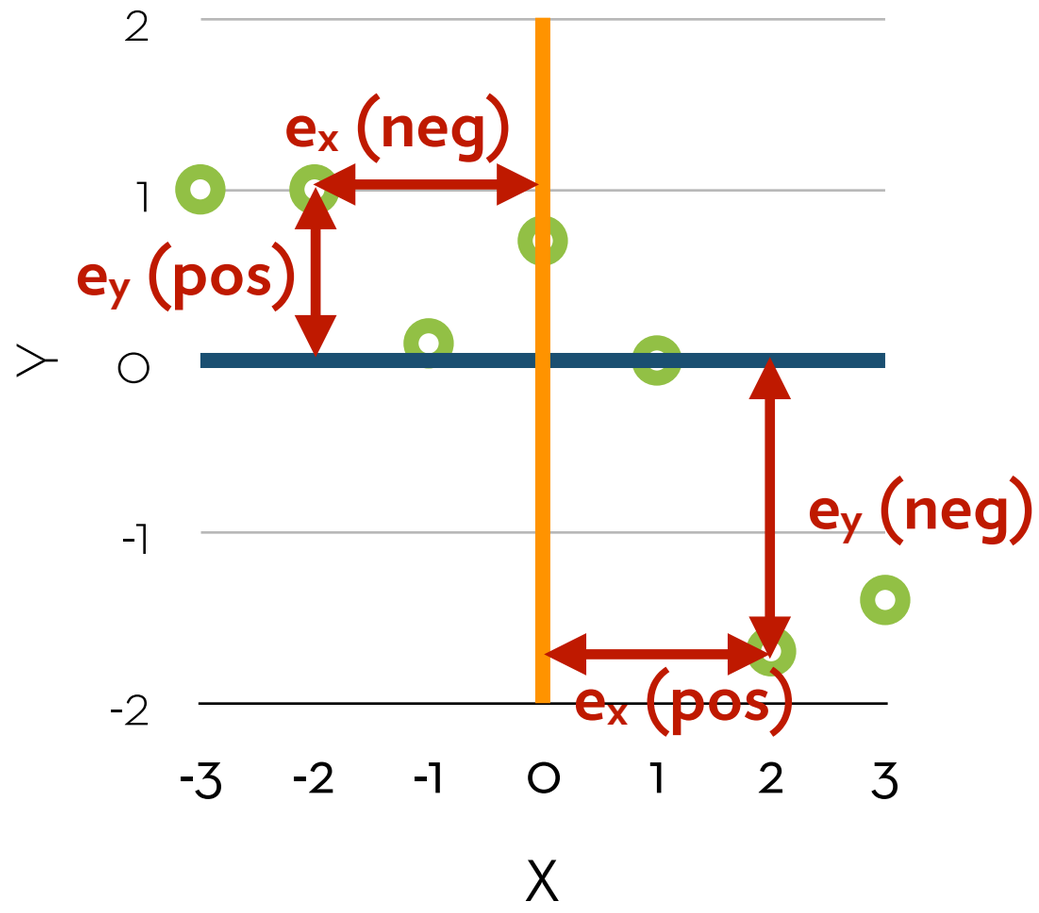


Covariance

Covariance is negative when:

The error in x is negative and the error in y is positive (-1*1 = -1)

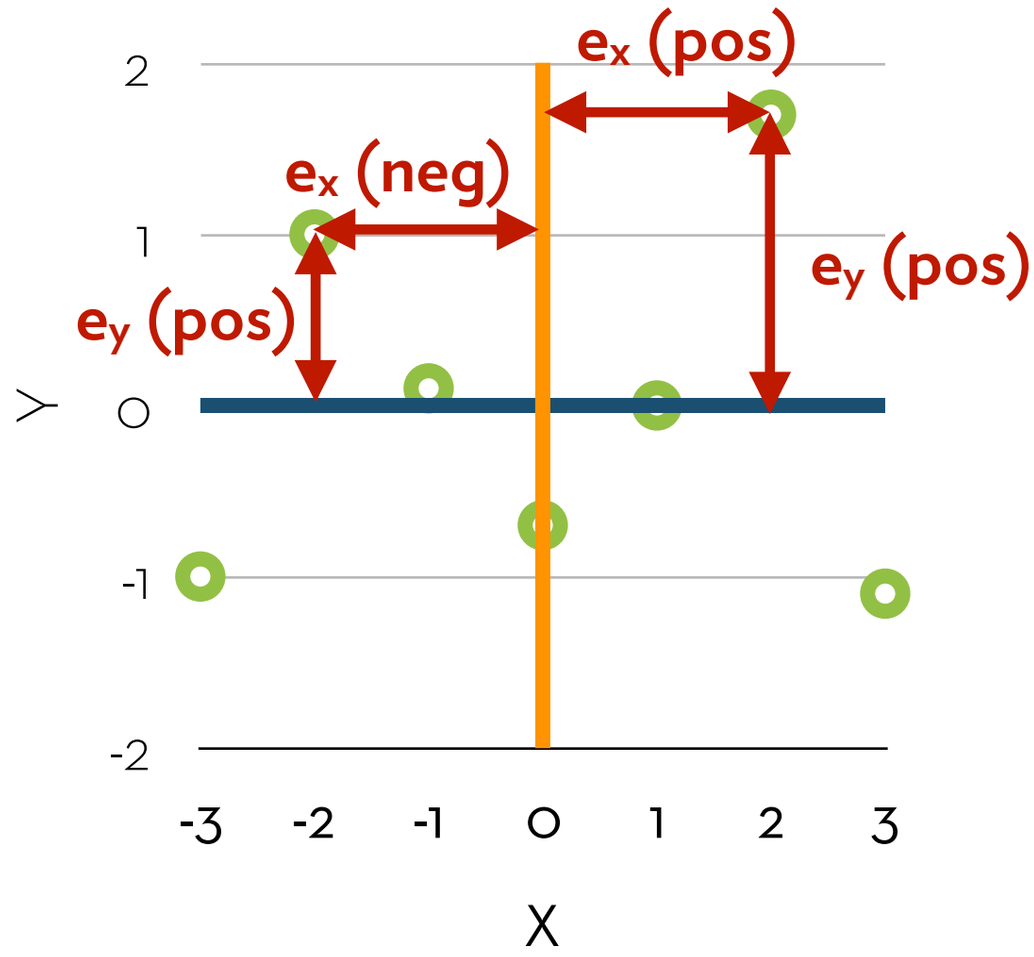
The error in x is positive and the error in y is negative (1*-1 = -1)



Covariance

Covariance is zero when:

Both of these situations occur almost equally



Covariance

Problem: covariance depends on the scale of measurement

If you multiply the scale of x or y by, say, 10, your covariance becomes 10 times as large!

Solution: standardization

We can standardize any deviation by dividing it by the **standard deviation** of the measure ($\sqrt{}$variance)

If we want to standardize the covariance, we divide by **both** the standard deviation of x and the standard deviation of y.

Correlation

The resulting metric is the **correlation coefficient**:

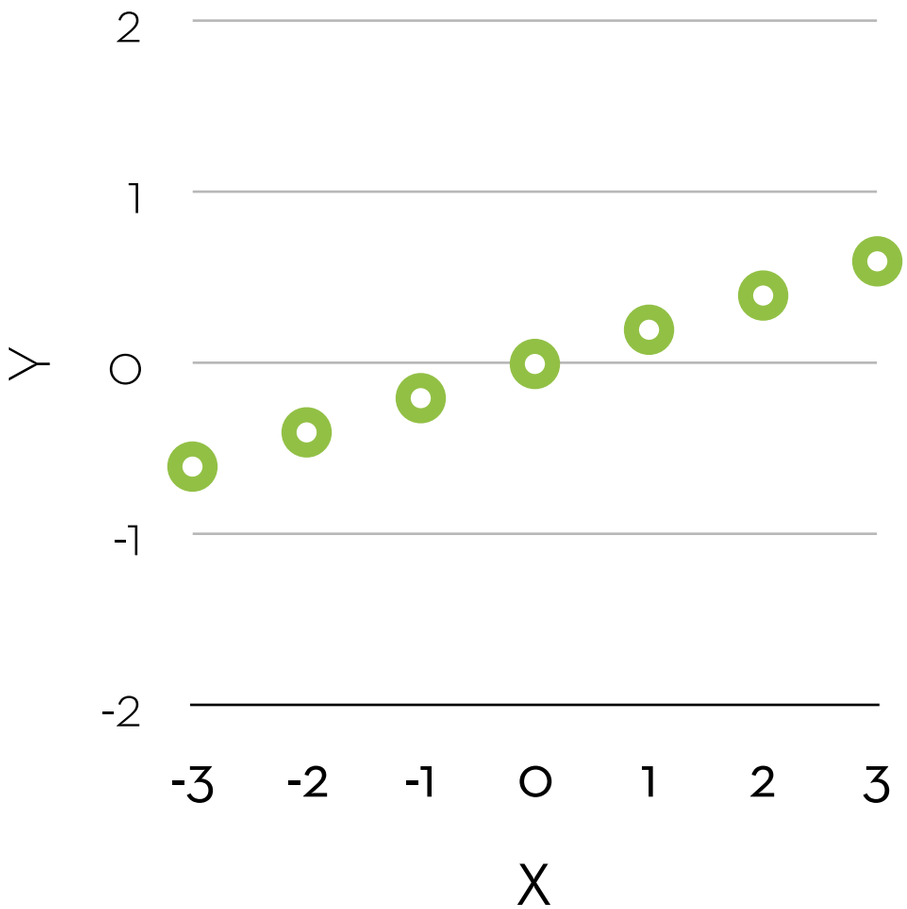
$r = cov(x,y)/s_x s_y = \sum(x_i - mean_x)(y_i - mean_y)/(N\text{-}1)s_x s_y$

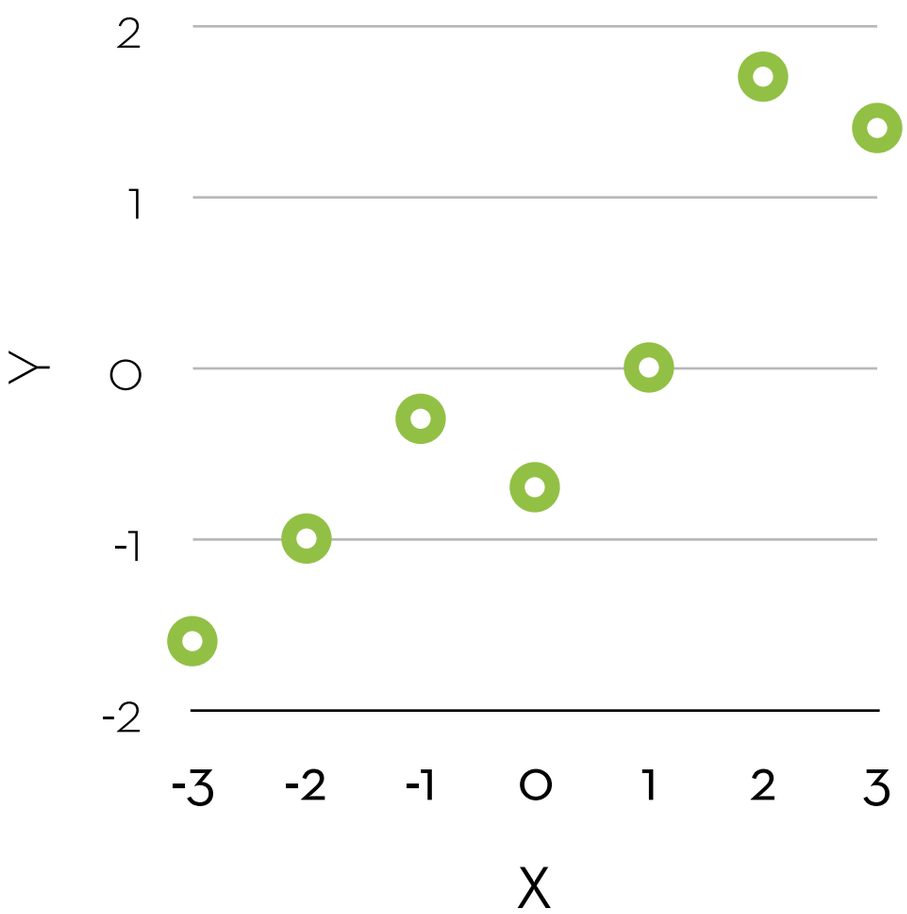
This measure falls between –1 and +1

- .1 is small
- .3 is medium
- .5 is large

Correlation

Which of these two graphs shows the strongest correlation?





Testing correlation

Can we test whether the correlation is significantly different from 0 (no correlation)? Yes!

Using z scores:

$z_r = 0.5*\ln((1+r)/(1-r))$, with SE = $1/\sqrt{(N-3)}$

Check z_r/SE against a z-table

Remember $z > 1.96$ means $p < .05$ (two-tailed)

Using t scores:

$t_r = r\sqrt{(N-2)}/\sqrt{(1-r^2)}$ with N-2 degrees of freedom

Creating a CI of r

Can we create a confidence interval of the correlation? Yes!

Using z scores:

upper z_r bound: $z_r + 1.96*SE$

lower z_r bound: $z_r - 1.96*SE$

Turn the upper and lower z_r back into upper and lower r:

$r = (e^{(2z_r)}-1)/(e^{(2z_r)}+1)$

Causation? No.

Correlation is not causation!

Just because two variables are correlated, doesn't mean that one causes the other

Remember: shoe size and intelligence are correlated!

Correlation in R

Because formulas suck, right?

Correlation in R

Three methods:

cor() - can do multiple correlations

cor.test() - can do p-values and CIs, but only on pairs

rcorr() - can do p-values and multiple correlations (but no CI)

Example

Dataset “Exam Anxiety.dat” -> set name to examData

Effect of exam stress on exam performance

Variables:

Code: participant id

Revise: hours spent revising

Exam: performance (%)

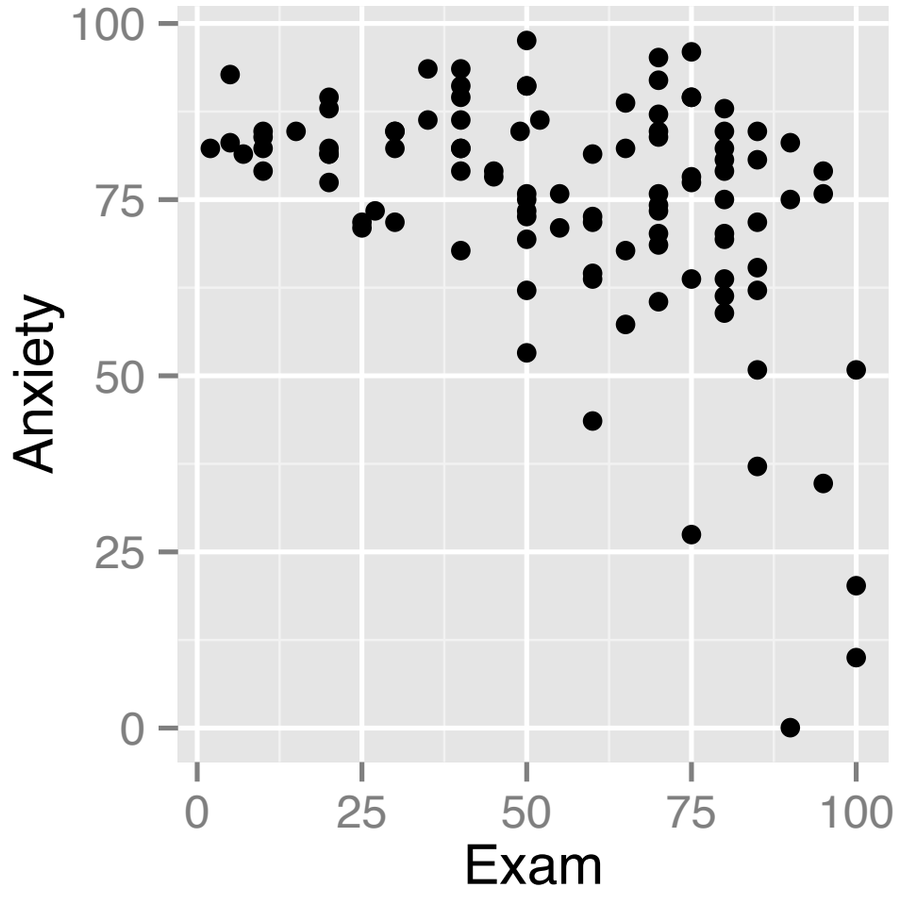
Anxiety: anxiety level (questionnaire score)

Gender: male/female

Example

Start with a scatterplot; x = Exam, y = Anxiety:

```
ggplot(examData, aes(Exam, Anxiety)) + geom_point()
```



Example

Run “cor” on Exam and Anxiety:

```
cor(examData$Exam, examData$Anxiety)
```

Let’s get p-values and CIs:

```
cor.test(examData$Exam, examData$Anxiety)
```

All combinations of variables:

```
cor(examData[,c(“Revise”,”Exam”,”Anxiety”])
```

All combinations, including p-values (load Hmisc package):

```
rcorr(as.matrix(examData[,c(“Revise”,”Exam”,”Anxiety”]))
```

R^2

If we square the correlation coefficient, we get the **coefficient of determination**, or R^2

Interpretation: the proportion of shared variation between x and y

> E.g. 19.4% of the variation in exam scores is shared by anxiety (so 80.6% of the variation is still unexplained!)

We are going to use this metric a lot!

Assumptions

Correlation is valid for any intervals data

Correlation **test** requires that both variables are normally distributed

Exception: technically, one variable can be binary

but then we are basically just conducting a t-test...

Advanced correlation

Robust methods, and partial correlations

Robust correlation

What if the data is not interval but ordinal? Or not normal?

- Use Spearman's rho
- Use Kendall's tau (generally better, especially if there are many of the same scores)
- Use bootstrapping (gives us confidence intervals)

Examples in R

Kendall Tau:

```
cor.test(examData$Exam, examData$Anxiety, method =
"kendall")
```

Examples in R

Bootstrapped Kendall Tau:

Create a function for running the bootstrap sample:

```
bootFun <- function(examSample,i) cor(examSample
$Exam[i], examSample$Anxiety[i], method = "kendall")
```

Run the bootstrap sample over the function:

```
bootResult <- boot(examData, bootFun, 2000)
```

Boot() creates 2000 random samples from examData, and runs them all on bootFun

Examples in R

Get output:

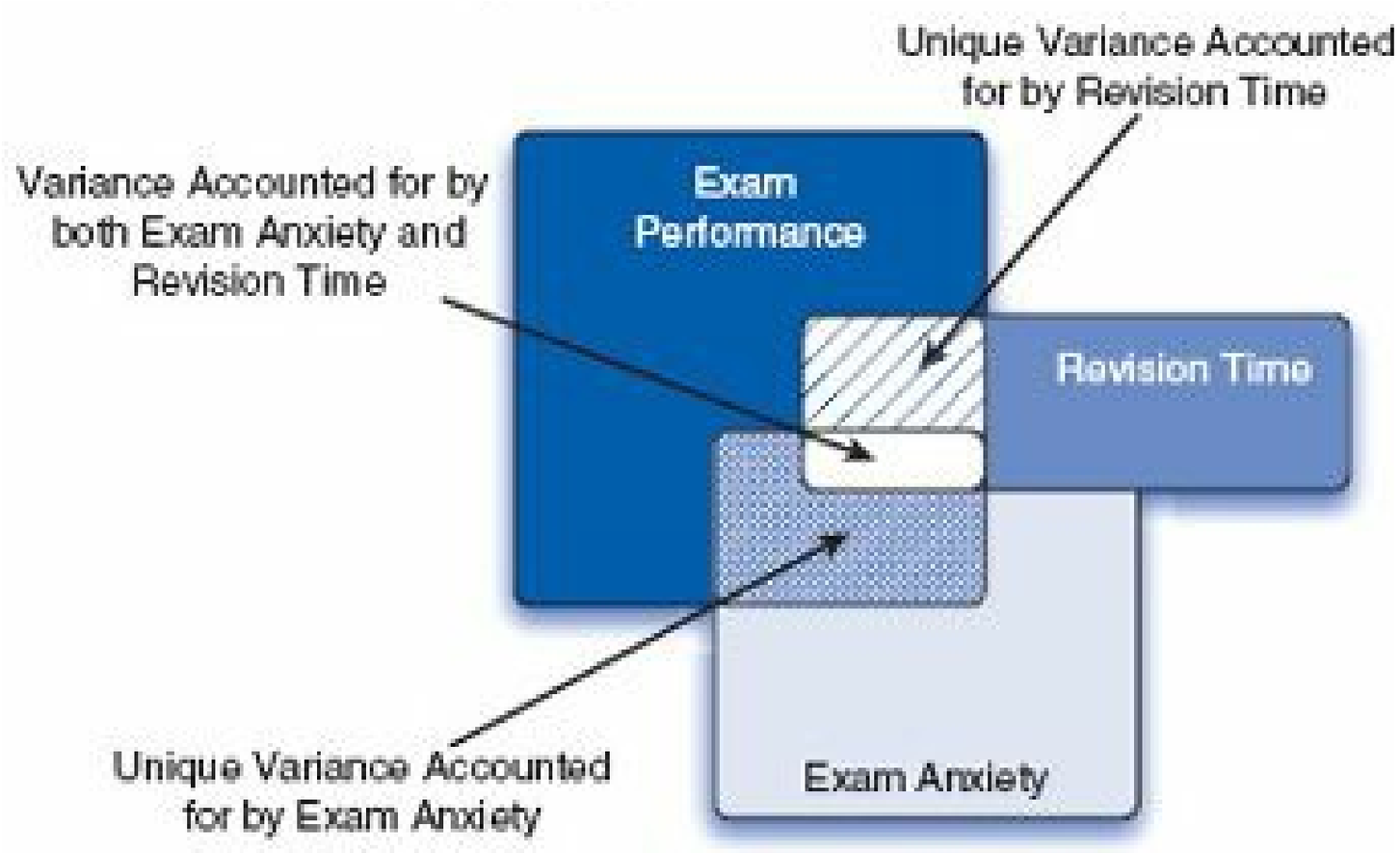
bootResult

This gives us a bias and a standard error

Finally, we can get a confidence interval:

boot.ci(bootResult)

Partial correlation



Partial correlation

Part of the correlation between Exam score and Anxiety can be explained by Revision time

> Can we get the correlation between Exam score and Anxiety, without the part that could be explained by (i.e. controlling for) Revision time?

Part of the correlation between Exam score and Revision time can be explained by Anxiety time

> Can we get the correlation between Exam score and Revision time, controlling for Anxiety?

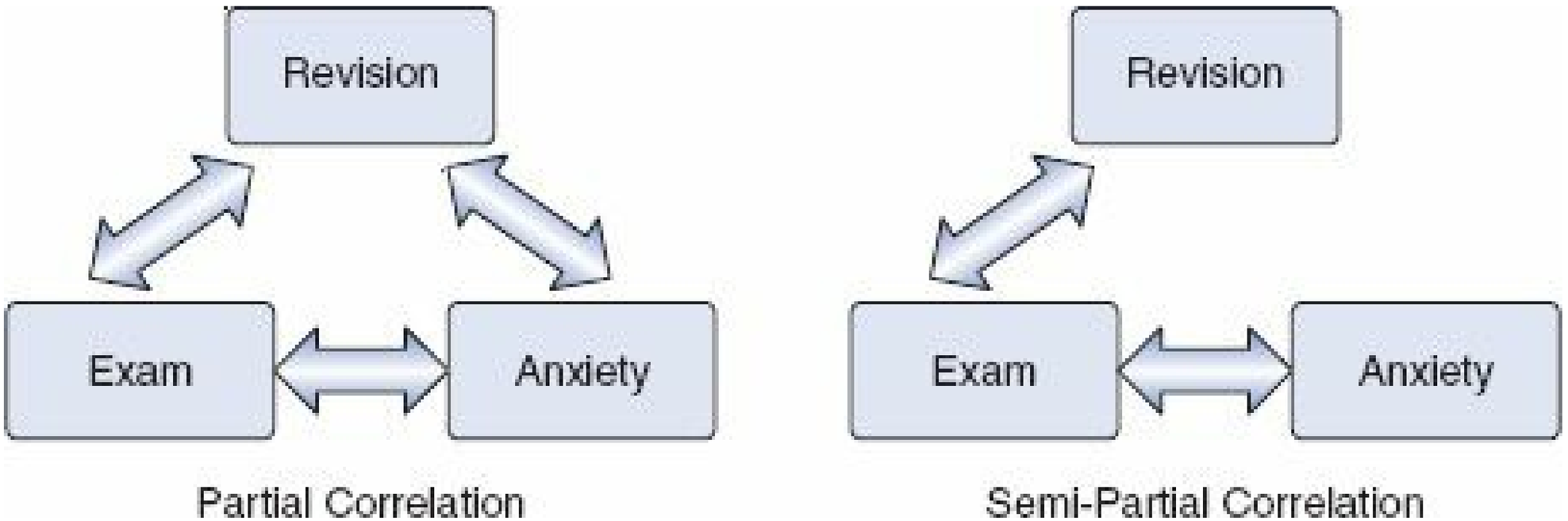
Partial correlation

Answer: Yes! With partial correlation!

In R:

- load “ggm” package
- pc <- pcor(c(“Exam”, “Anxiety”, “Revise”), var(examData))
- pc (gives the partial correlation)
- pcor.test(pc, 1, 103) (gives the p-value)

Part correlation



More on this later...

Reporting

"Exam performance was significantly correlated with exam anxiety, r = -.44, p < .001"

"There was a significant relationship between exam performance and time spent revising, r = .40, p < .001"

"Exam anxiety was significantly related to the time spent revising, r = -.71, p < .001"

Additional stuff

Things not covered:

- Comparing correlations (independent or dependent)
- Biserial and point-biserial correlations

These will **not** be on the exam / assignments

They may help you understand some of our future tests, though!

“It is the mark of a truly intelligent person
to be moved by statistics.”

T H A N K S !

George Bernard Shaw