Online Academic Data Analysis Bootcamp Using Open-Access Program R

Correlation

Peninah Karomo
Department of Geography & Environmental Studies

James Orwa
(Biostatistician/Instructor) Ph.D. Student UoN, Master of Statistics, Hasselt Belgium

Dr. Patrick Njage
Technical University of Denmark and
Academic Data Analysts (https://www.academicdataanalysts.org/)

Outline

Measuring Relationships
- Scatterplots
- Covariance
- Pearson's Correlation Coefficient

Nonparametric measures
- Spearman's Rho
- Kendall's Tau

Interpreting Correlations
- Causality

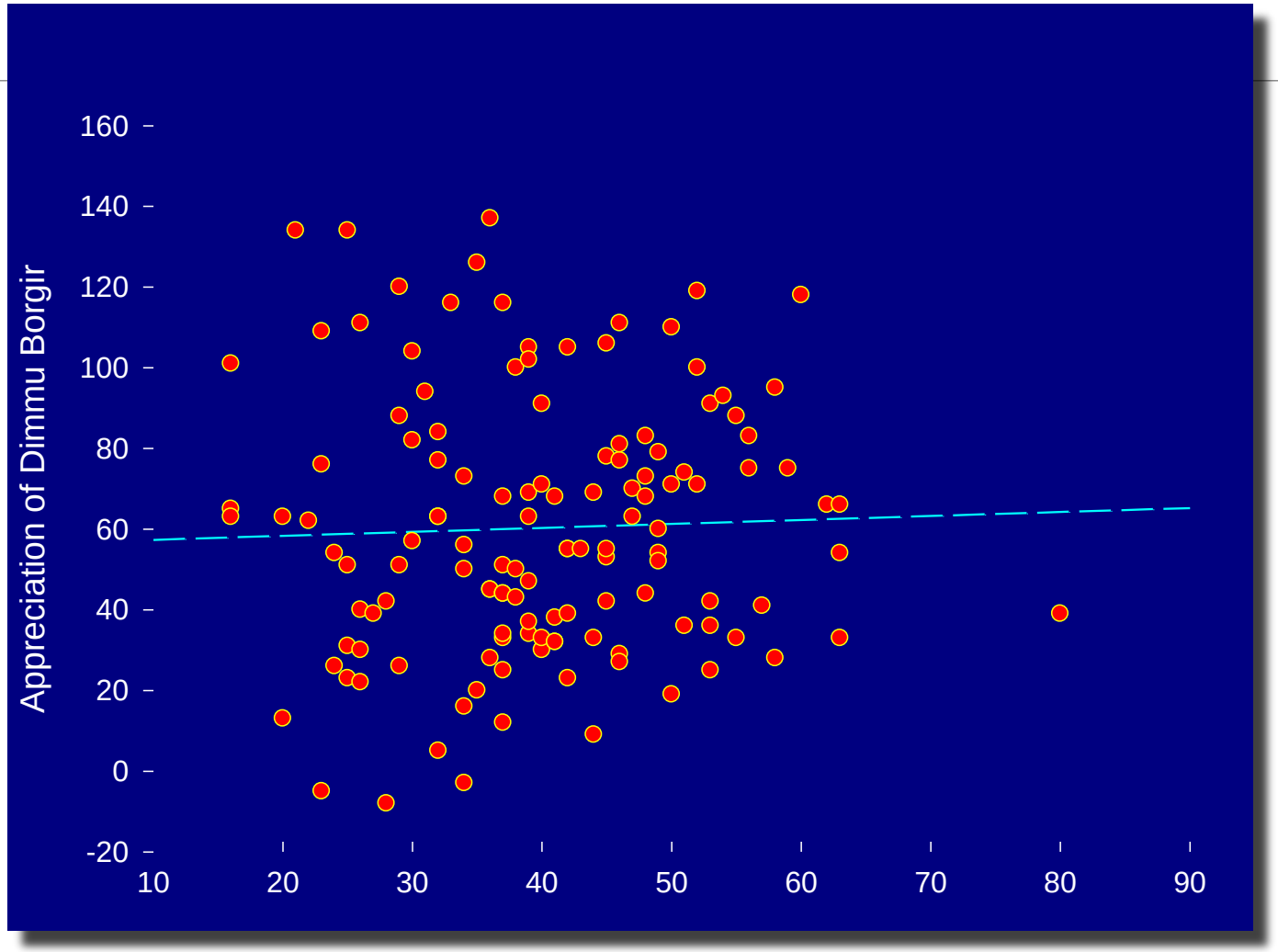
Partial Correlations

What is a Correlation?

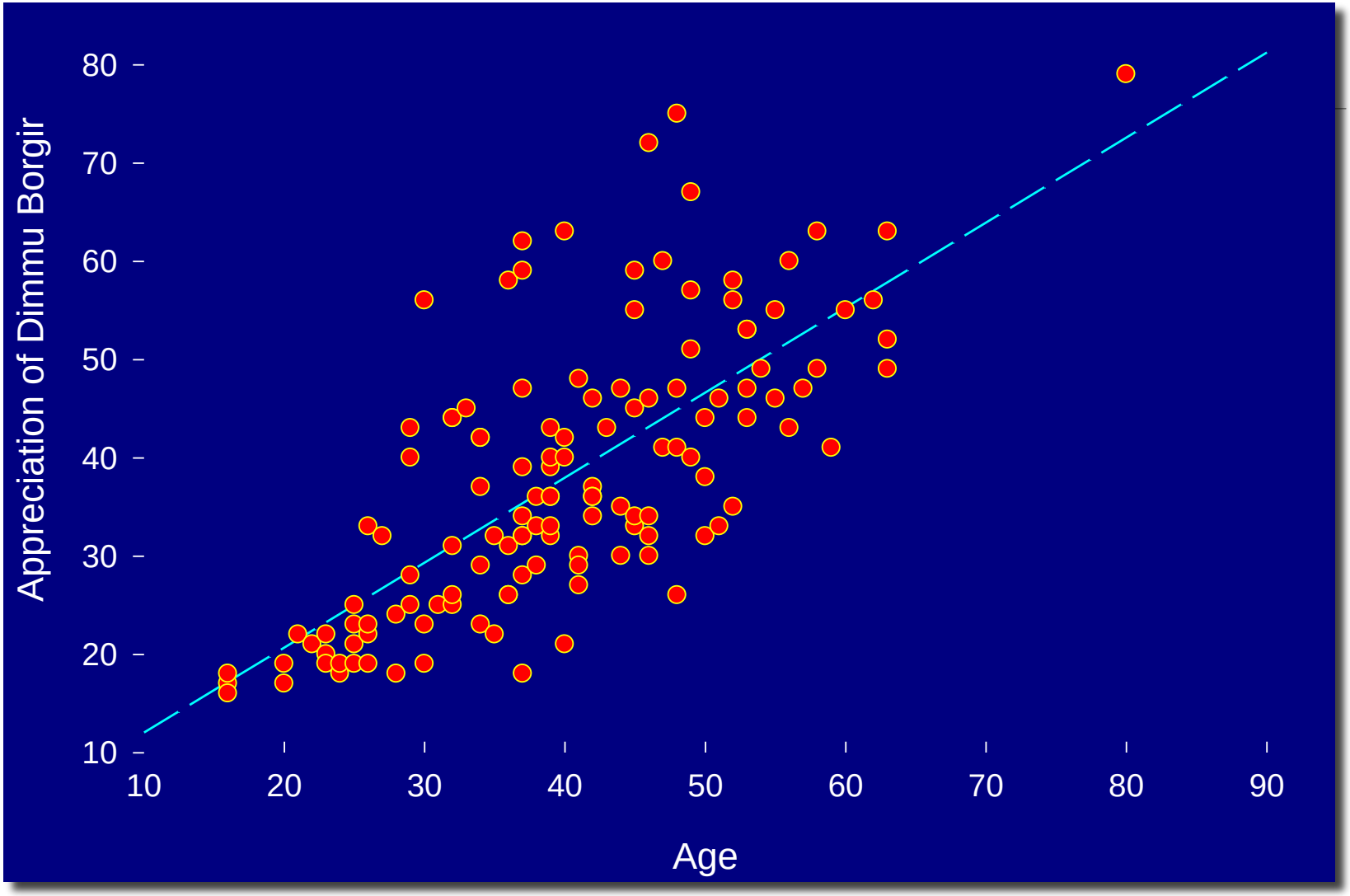
It is a way of measuring the extent to which two variables are related.

It measures the pattern of responses across variables.

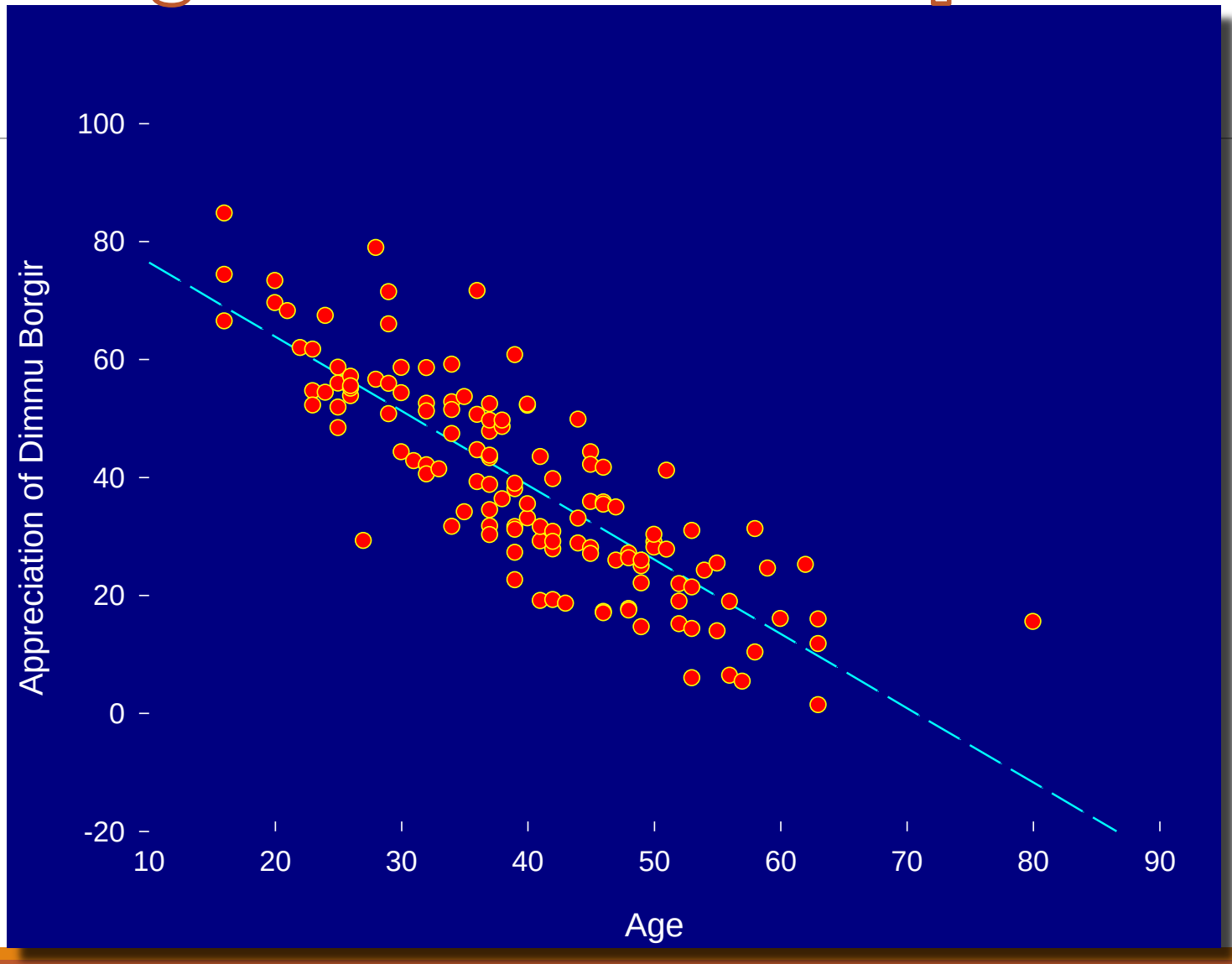
Very Small Relationship



Positive Relationship



Negative Relationship



Measuring Relationships

We need to see whether as one variable increases, the other increases, decreases or stays the same.

This can be done by calculating the Covariance.

- We look at how much each score deviates from the mean.
- If both variables deviate from the mean by the same amount, they are likely to be related.

Modeling Relationships

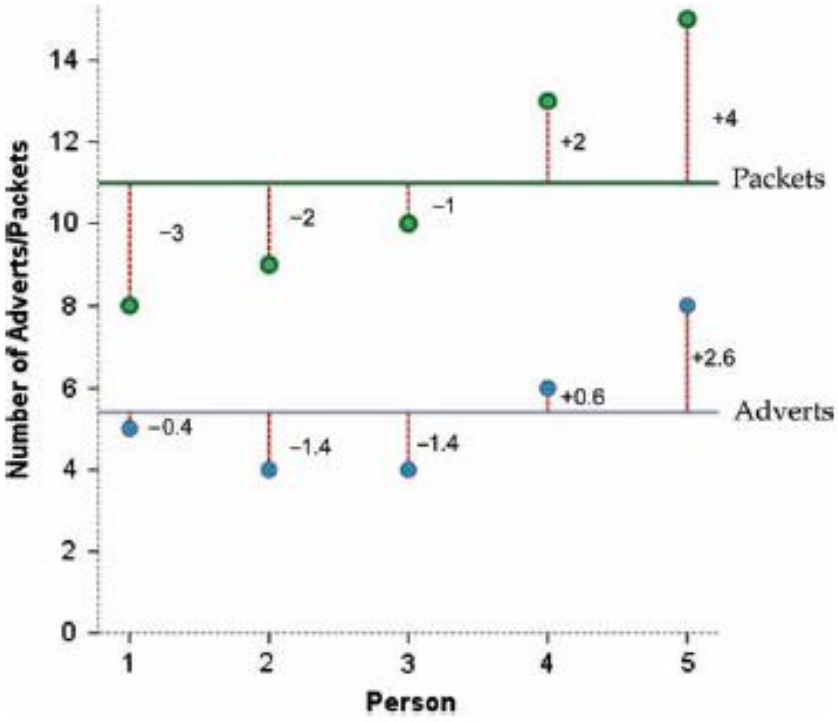
First, look at some scatterplots of the measured variables.

$$\text{Outcome}_i = (\text{model}) + \text{error}_i$$

$$\text{Outcome}_i = (bX_i) + \text{error}_i$$

Adverts watched versus toffee purchases

Participant:	*1*	*2*	*3*	*4*	*5*	*Mean*	*s*
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92



Grey line: average number of packets bought

Blue line: average number of adverts watched

Vertical lines: differences between the observed values and the mean of the relevant variable (called deviations)

Notes

Very similar pattern of deviations for both variables.

First three participants: observed values are below the mean for both variables

Last two people: observed values are above the mean for both variables.

This pattern is indicative of a potential relationship between the two variables - if a person's score is below the mean for one variable then their score for the other will also be below the mean).

Revision: variance

- The variance tells us by how much scores deviate from the mean for a single variable.

- It is closely linked to the sum of squares.

- Covariance is similar – it tells is by how much scores on two variables differ from their respective means.

Revision: variance

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$= \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$\bar{x}$: the mean of the sample

x_i: the data point in question

N: the number of observations

Covariance

- Calculate the error between the mean and each subject's score for the first variable (*x*).

- Calculate the error between the mean and their score for the second variable (*y*).

- Multiply these error values.

- Add these values and you get the cross product deviations.

- The covariance is the average cross-product deviations:

Covariance equation

Note: similar to the equation for variance, except that **instead of squaring** the differences, we multiply them by the corresponding **difference of the second variable**

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Example: Adverts watched and toffee purchases

Participant:	1	2	3	4	5	Mean	s
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

$$\begin{aligned}
\text{cov}(x,y) &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1} \\
&= \frac{(-0.4)(-3)+(-1.4)(-2)+(-1.4)(-1)+(0.6)(2)+(2.6)(4)}{4} \\
&= \frac{1.2+2.8+1.4+1.2+10.4}{4} \\
&= \frac{17}{4} \\
&= 4.25
\end{aligned}$$

Problems with Covariance

It **depends upon the units of measurement.**

- E.g. The Covariance of two variables measured in Miles might be 4.25, but if the same scores are converted to Km, the Covariance is 11.

One solution: standardise it!

- **Divide by the standard deviations of both variables**.

The standardised version of Covariance is known as the Correlation coefficient.

- It is relatively unaffected by units of measurement.

The Correlation Coefficient

$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

Correlation coefficient: express the deviation from the mean for a participant in standard units by dividing the observed deviation by the standard deviation.

s_x is the standard deviation of the first variable and s_y is the standard deviation of the second variable (all other letters are the same as in the equation defining covariance)

The Correlation Coefficient

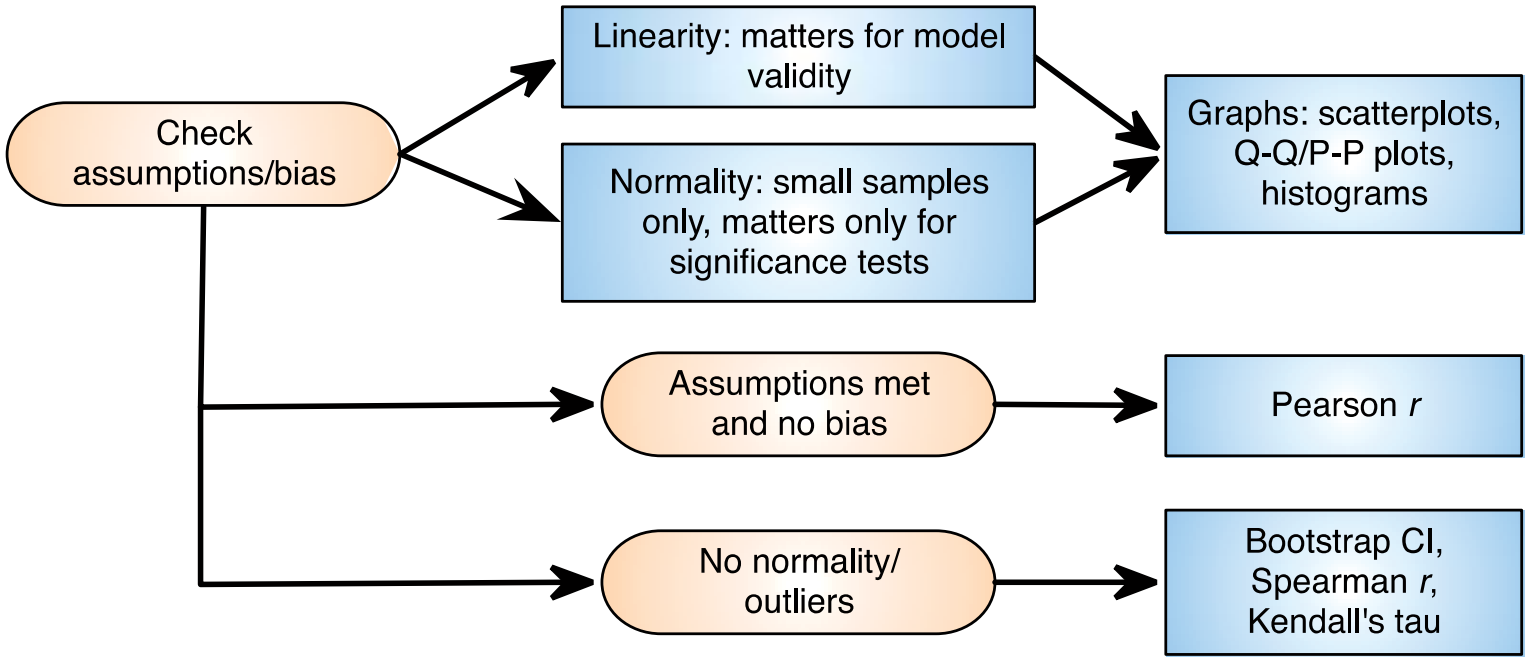
$$r = \frac{Cov_{xy}}{s_x s_y}$$

$$= \frac{4.25}{1.67 \times 2.92}$$

$$= .87$$

Conducting Correlation Analysis

The general process for conducting correlation analysis



Correlation Example

```
> adverts      <-c (5,4,4,6,8)
> packets      <-c (8,9,10,13,15)
> advertData       <-data.frame(adverts,packets)

> with(advertData,
cor.test(adverts, packets,  alternative="two.sided",method="pearson"))
```

Pearson's product-moment correlation data:
adverts and packets t = 82191000, df = 3, p-value <
2.2e-16

alternative hypothesis:
correlation is not equal to 0

Correlation: Example

Anxiety and Exam Performance

Participants:

- 103 students

Measures

- Time spent revising (hours)
- Exam performance (%)
- Exam Anxiety (the EAQ, score out of 100)
- Gender

Correlation: Example

- Exam Anxiety Dataset

- Variables
  - 4 continuous: Code, Revise, Exam, Anxiety
  - 1 categorical: Gender (Male / Female)

Code	Revise	Exam	Anxiety	Gender
1	4	40	86.298	Male
2	11	65	88.716	Female
3	27	80	70.178	Male
4	53	80	61.312	Male
5	4	40	89.522	Male
6	22	70	60.506	Female

Bivariate Correlation Test

- Two Correlation Approaches:
  - Test: 2 variables (single bi-variate test)
  - Matrix: > 2 variables (multiple pair-wise tests)

- Three Metrics: Pearson r, Spearman Rank, Kendall's Tau

- Recommendation:

  Use correlation tests (cor.test) to test the significance of correlations (p values)

- 3 CorrelationTypes:
  - Pearson r
  - Kendall's Tau
  - Spearman Rank

- 2 Alternatives:
  - Non-Directional (2-sided)
  - Directional (1-sided) Positive, Negative

Correlation in r: How to

```
Rcmdr>
with(exam, cor.test(Anxiety, Exam,  alternative="two.sided",
method="pearson"))

Pearson's product-moment correlation data:
Anxiety and Exam
t = -4.938, df = 101
p-value = 0.000003128

alternative hypothesis:
true correlation is not equal to 0

95% confidence interval:
-0.5846244 -0.2705591

sample estimates: cor -0.4409934
```

Sample size? 101 +2 = 103

Significant ? YES (P < 0.05)

1 or 2 tailed test ? 2

Significant? YES (no 0 overlap)

Correlation in r: How to

- Three functions that can be used to compute basic correlation coefficients:

  cor() AND cor.test()

Function	Pearson	Spearman	Kendall	*p*-values	CI	Multiple Correlations?
cor()	✓	✓	✓			✓
cor.test()	✓	✓	✓	✓	✓	

NOTE:

rcorr() function needs "Hmisc" package

Correlation in r: How to

```
Rcmdr>  cor(exam[,c("Anxiety","Exam","Revise")],
use="complete")
```

	Anxiety	Exam	Revise
Anxiety	1.0000000	-0.4409934	-0.7092493
Exam	-0.4409934	1.0000000	0.3967207
Revise	-0.7092493	0.3967207	1.0000000

NOTE:

Correlation matrix does not report p values, no 95% C.I., and no sample size information

Reporting Correlation Results

- Report 4 pieces of information = test, r, df, p value (*)

	Exam Performance	Exam Anxiety	Revision Time
Exam Performance	1	−.44***	.40***
Exam Anxiety	101	1	−.71***
Revision Time	101	101	1

Ns = not significant ($p > .05$), *$p < .05$, ** $p < .01$, *** $p < .001$

(df = n - 2)

- **Other Statistics to Report:**

  r^2 (Coefficient of determination)

  Squaring the *r* gives you the proportion of variance in one variable shared by the other

Reporting Correlation Results

- Report 4 pieces of information = test, r, df, p value.

Based on three pair-wise Pearson correlations (df = 101), we documented the following significant correlations:

Exam performance was negatively correlated with the exam anxiety ($r = -0.44$, $p < 0.001$), and positively correlated with the time spent revising, ($r = +0.40$, $p < 0.001$). The time spent revising was also negatively correlated with the exam anxiety, ($r = -0.71$, $p < .001$).

Reporting the Results

TABLE 7.2 An example of reporting a table of correlations

	Exam Performance	*Exam Anxiety*	*Revision Time*
Exam Performance	1	−.44*** [−.564, −.301]	.40*** [.245,.524]
Exam Anxiety	103	1	−.71*** [−.863, −.492]
Revision Time	103	103	1

ns = not significant ($p > .05$), * $p < .05$, ** $p < .01$, *** $p < .001$. BCa bootstrap 95% CIs reported in brackets.

Things to know about the Correlation

It varies between -1 and +1

- 0 = no relationship

It is an effect size

- ±.1 = small effect
- ±.3 = medium effect
- ±.5 = large effect

Coefficient of determination, r^2

- By squaring the value of *r* you get the proportion of variance in one variable shared by the other.

Experimental Research Methods

Cause and Effect (Hume, 1748)

1. Cause and effect must occur close together in time (contiguity);
2. The cause must occur before an effect does;
3. The effect should never occur without the presence of the cause.

Confounding variables: the '*Tertium Quid*'

- A variable (that we may or may not have measured) other than the predictor variables that potentially affects an outcome variable.
- E.g. The relationship between breast implants and suicide is confounded by self-esteem.

Ruling out confounds (Mill, 1865)

- An effect should be present when the cause is present and when the cause is absent the effect should be absent also.
- Control conditions: the cause is absent.

Correlation and Causality

The third-variable problem:

- in any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.

Direction of causality:

- Correlation coefficients say nothing about which variable causes the other to change

Nonparametric Correlation

Spearman's Rho

- Pearson's correlation on the ranked data

Kendall's Tau

- Better than Spearman's for small samples

World's best Liar Competition

- 68 contestants
- Measures
  - Where they were placed in the competition (first, second, third, etc.)
  - Creativity questionnaire (maximum score 60)

Try data transformation

Use non-parametric test:

- Spearman Rank (rho)
- Kendall's Tau

Recommendations:

- Spearman's rho (Pearson's correlation on ranked data)
  - Better for large datasets
  - No problem with tied ranks
- Kendall's tau
  - Better for small samples sizes
  - Has a problem with tied ranks

Formulation

Kendall rank correlation: A non-parametric test that does not make any assumptions about the distributions - unlike the Pearson's correlation.

Kendall rank index, Tau:

Kendall's tau, τ, is another non-parametric correlation and it should be used rather than Spearman's coefficient when you have a small data set with a large number of tied ranks.

Nonparametric Correlations

Example: World's Biggest Liar competition:

- File: TheBiggestLiar.xlsx
- N = 68 contestants
- Measures:
  - Where they were placed in the competition (first, second, third, etc.)
  - Creativity questionnaire (max score 60)

Nonparametric Correlation: Sperman's Rank

- To calculate correlation, use command:

```
cor(liarData$Position,  liarData$Creativity, method = "spearman")
```

- Output of this command is: [1] -0.3732184
- To get significance value, use command:

```
cor.test(liarData$Position,  liarData$Creativity, method = "spearman")
```

- Output of this command is:

Spearman's rank correlation rho data:

liarData$Position S = and liarData$Creativity

71948, p-value = 0.00172

alternative hypothesis: true rho is not equal to 0 sample estimates: rho -0.3732184

Nonparametric Correlations: Kendall's Tau

- To calculate correlation, use command:

cor(liarData$Position, liarData$Creativity, method = "kendall")

- Output of this command is: [1] -0.3002413
- To get significance value, use command:

cor.test(liarData$Position, liarData$Creativity, method = "kendall")

- Output of this command is:

Kendall's rank correlation tau data:

liarData$Position and liarData$Creativity

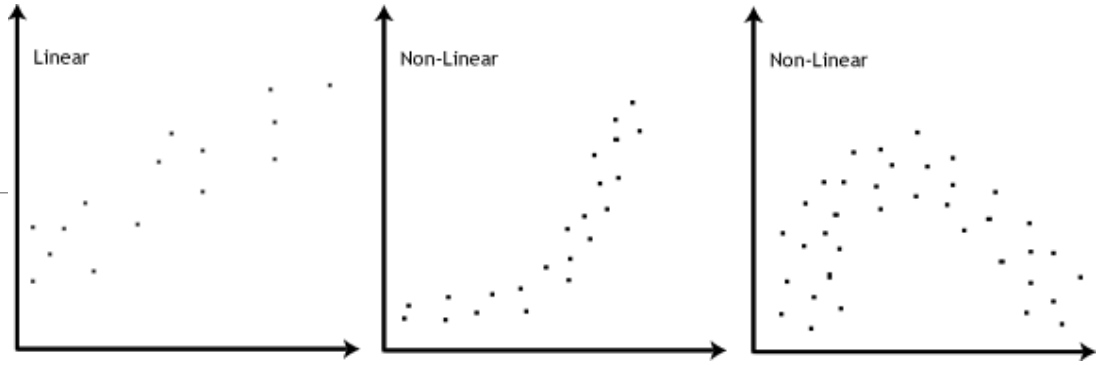
z = -3.2252, p-value = 0.001259

alternative hypothesis: true tau is not equal to 0

sample estimates: tau -0.3002413

Nonparametric Correlation - Hints

Relationship between variables NOT linear



Only one coefficient: sign / strength of association

Coefficient must always be in the range from –1 to 1

Which Nonparametric Correlation to Use

Kendall Tau works better with smaller samples

Spearman Rank better when there are tied data

Causality

- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change

- The third-variable problem:
  - In any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting results.

NOTE: Demonstrating causation requires controlled experiments (with controls)

Significance

Degrees of freedom influence the p value (significance)

Beware:
With larger sample sizes, it is easier to have a significant correlation

df = n -2				
Level of Significance (p) for Two-Tailed Test	**.10**	**.05**	**.02**	**.01**
df				
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708

df = n -2				
Level of Significance (p) for Two-Tailed Test	**.10**	**.05**	**.02**	**.01**
df				
25	.323	.381	.445	.487
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.303
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254

(Point-)biserial correlation

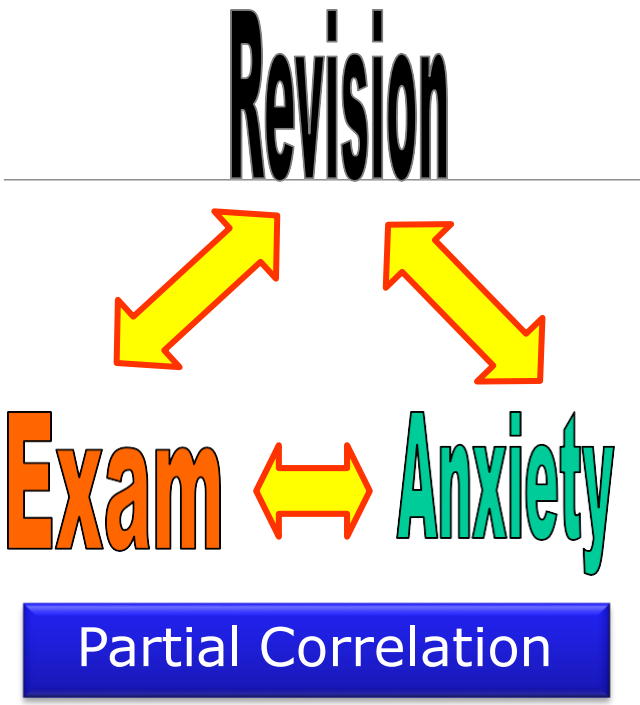
Point-biserial correlation, r_{pb}: relationship between

- a continuous variable and
- a variable that is a discrete dichotomy (no underlying continuum)

Biserial correlation, r_b: relationship between

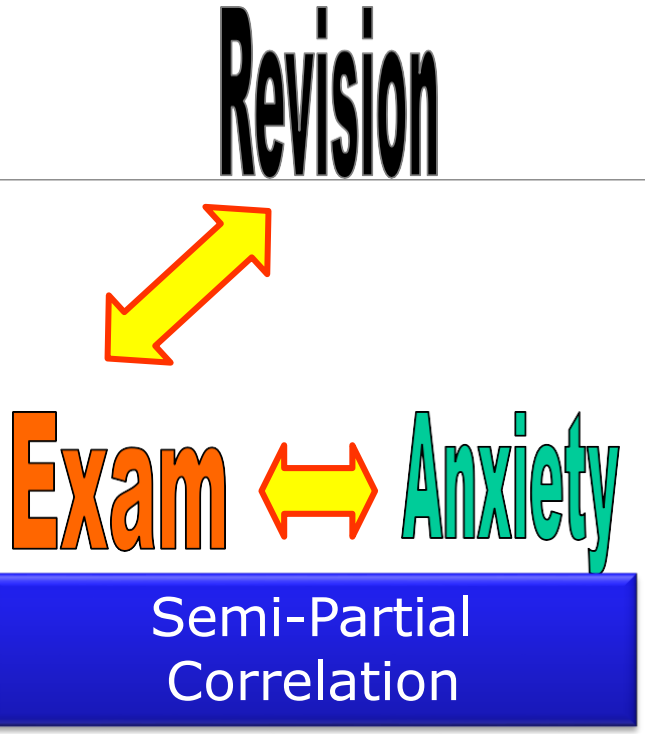
- a continuous variable and
- a variable that is a continuous dichotomy (continuum underlying the two categories)

Partial and Semi-partial Correlations



Partial Correlation

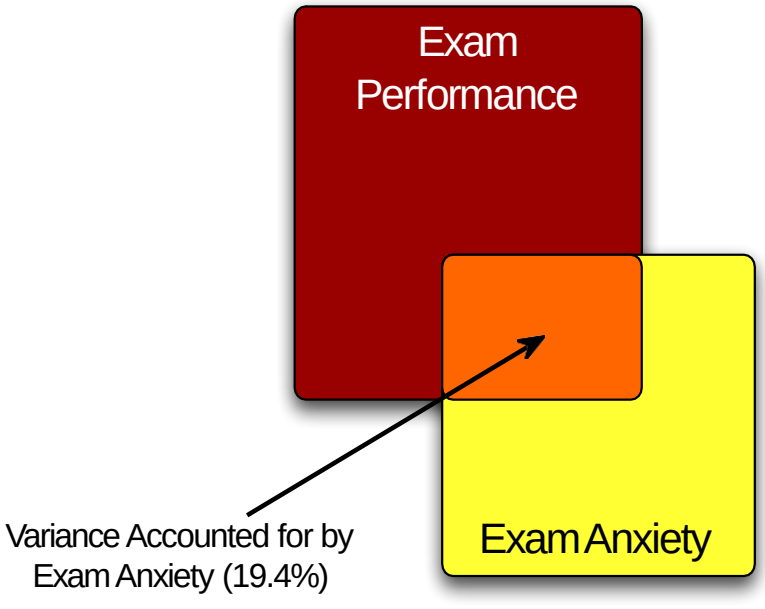
Measures relationship between two variables, controlling for effect that a third variable has on both of them.

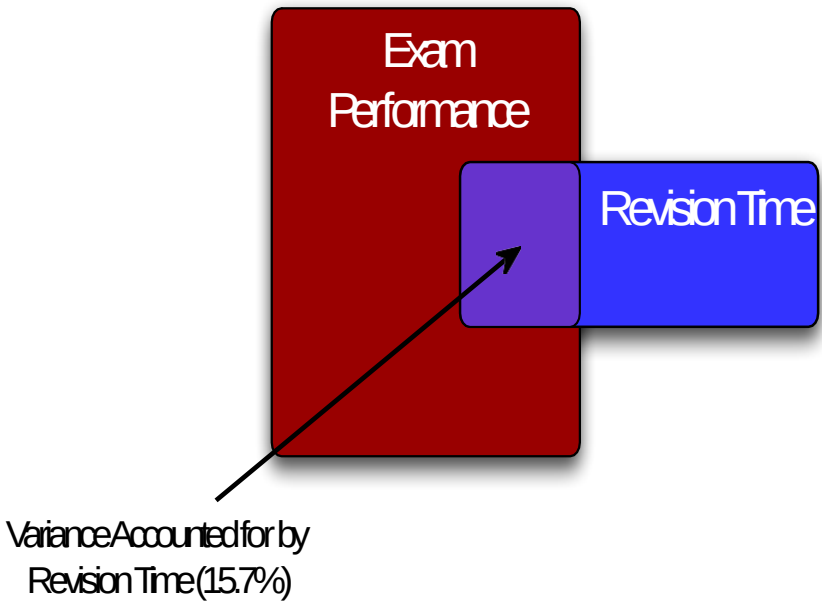


Semi-Partial Correlation

Measures relationship between two variables controlling for effect that a third variable has on one of them.

Shared Variance

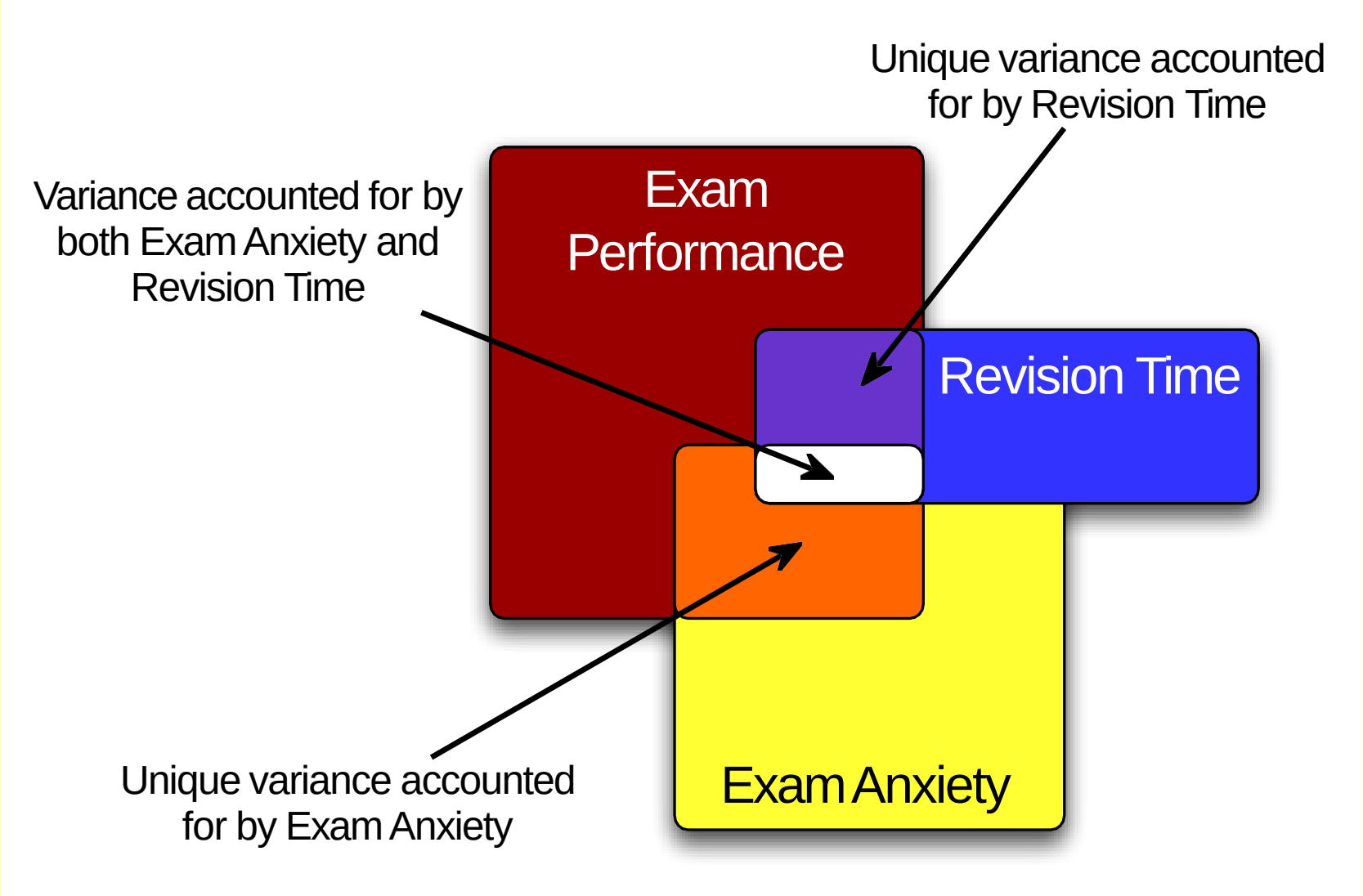




Exam ⟺ Anxiety

Revision ⟺ Anxiety

Shared Variance



Partial Correlations in r

NOTE: Requires installing package "ggm"

- The general form of *pcor()* is:
  pcor(c("var1", "var2", "control1", "control2" etc.), var(dataframe))

Partial Correlations in r

```
> library("ggm")                                   (use ggm package)

> examData2 <- examData[, c("Exam", "Anxiety",
"Revise")]                                   (made a new dataframe)

>pc <- pcor(c("Exam", "Anxiety", "Revise"), var(examData2))
    (create pcor object)

> pc [1] -0.2466658                  (pc = partial correlation)

> pc^2 [1] 0.06084403                (pc^2 = r squared)
```

Partial Correlations in r

- The general form of *pcor.test()* is:
  pcor(pcor object, number of control variables, sample size)

- Basically, enter an object you created with *pcor()* (or you can put the *pcor()* command directly into the function):

  pcor.test(pc, 1, 103)

Partial Correlations in r

- pc <- pcor(c("Exam", "Anxiety", "Revise"), var(examData2)) (create pcor object)

> pcor.test(pc, 1, 103)

$tval [1] -2.545307 (t test statistic)
$df [1] 100 (degrees freedom)
$pvalue [1] 0.01244581 (p value)

Interpreting Partial Correlations

```
Rcmdr>  cor(exam[,c("Anxiety","Exam","Revise")],
use="complete")
```

	Anxiety	Exam	Revise
Anxiety	1.0000000	-0.4409934	-0.7092493
Exam	-0.4409934	1.0000000	0.3967207
Revise	-0.7092493	0.3967207	1.0000000

Anxiety & Exam Correlation controlling for Revise

r = -0.2466658

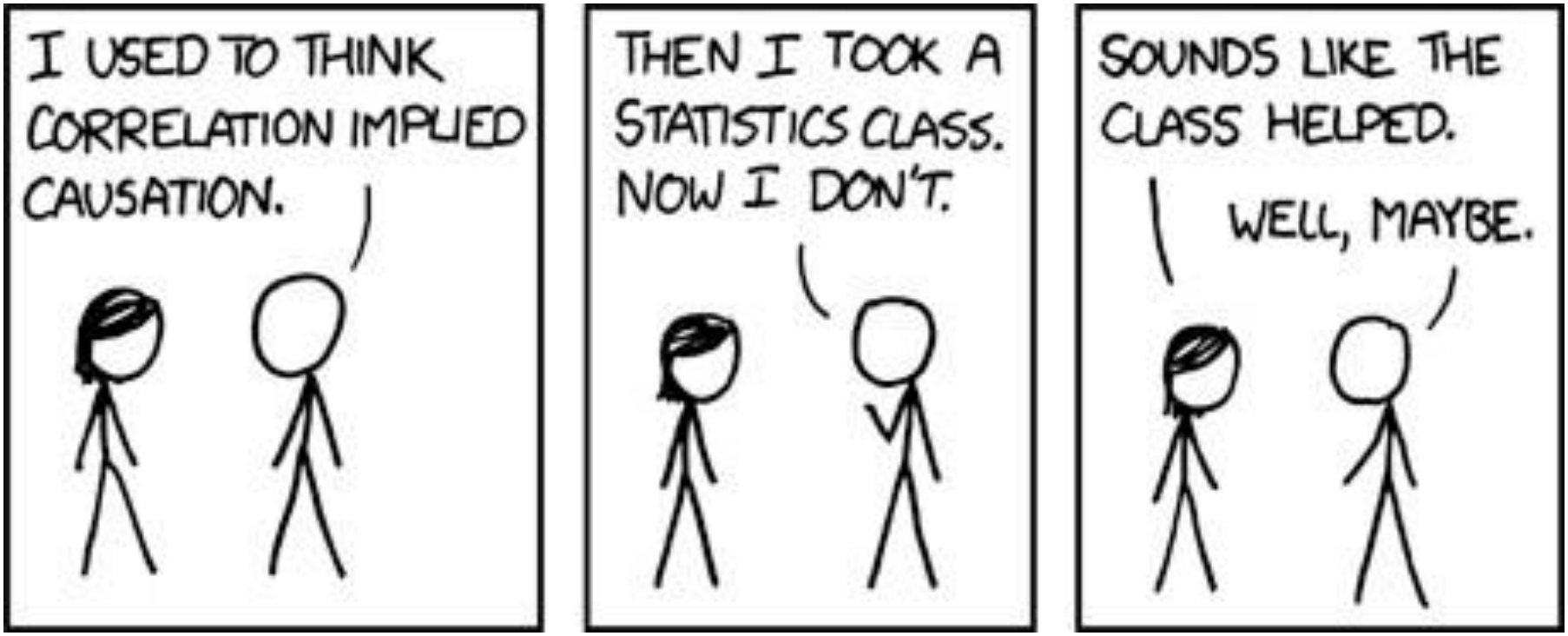
Why?

Due to effect of Revise on other two variables

Correlation Interpretation

- Direction of causality:
  - Determining cause-effect relationships requires controlled experiments
- The third-variable problem:
  - Partial correlations with other variables
  - Beware of unmeasured variables

Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.