Correlation

Vanessa LoBue

Jamil Bhanji

with a little help from Andy Field

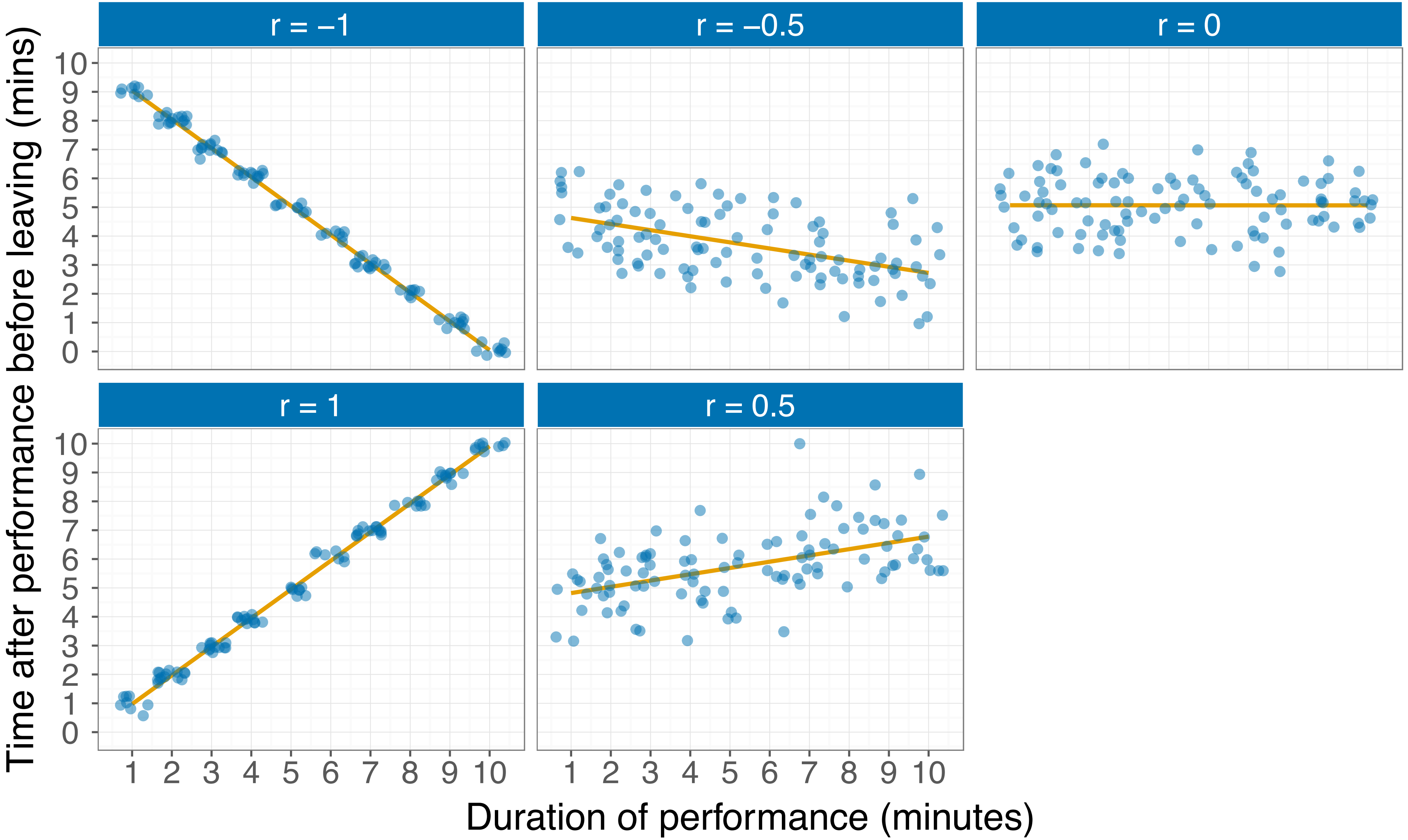
Aims

- 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



r = −1
r = −0.5
r = 0
r = 1
r = 0.5
Time after performance before leaving (mins)
Duration of performance (minutes)
0
1
2
3
4
5
6
7
8
9
10

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

Variance (review)

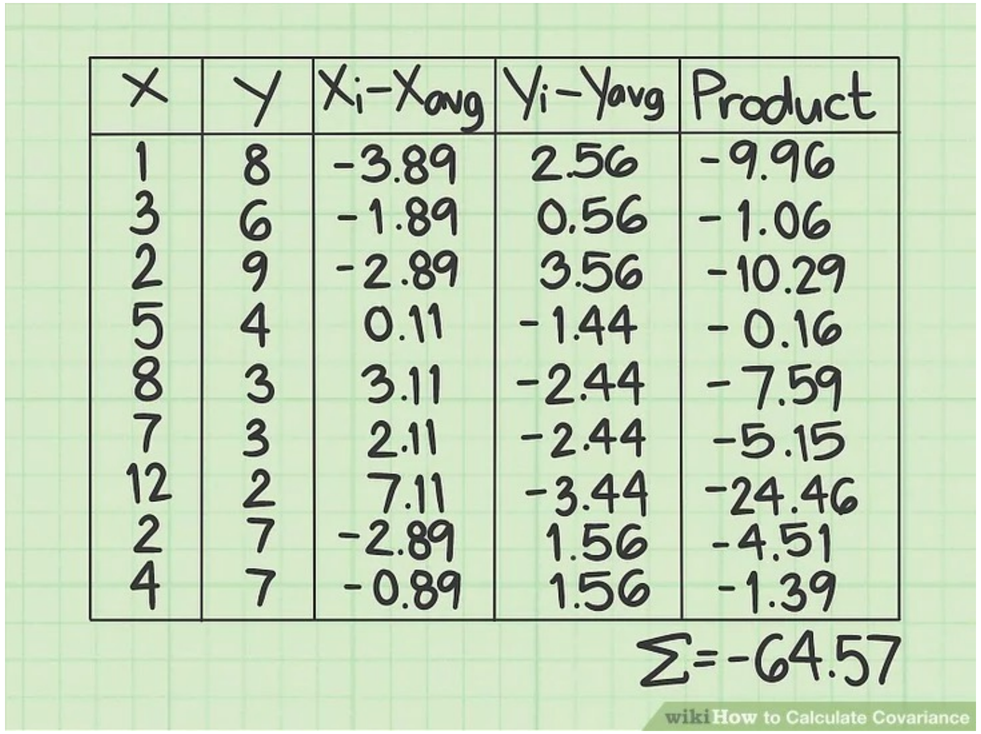
- The variance tells us by how much scores deviate from the mean for a single variable

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

Covariance

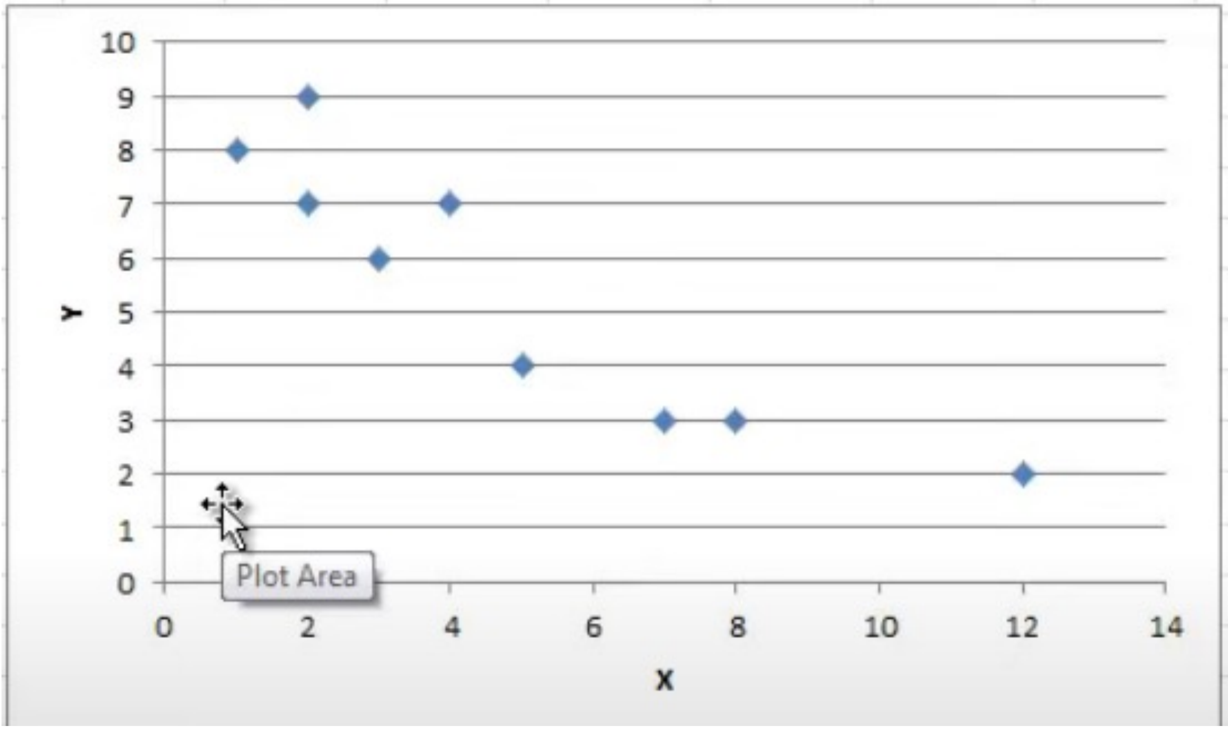
- Calculate the error between the mean and each subject's score for the first variable (*x*).
- Calculate the error between the mean and their score for the second variable (*y*).
- Multiply these error values.
- Add these values and you get the cross product deviations.
- The covariance is the average cross-product deviations

Covariance



X	Y	$X_i - X_{avg}$	$Y_i - Y_{avg}$	Product
1	8	-3.89	2.56	-9.96
3	6	-1.89	0.56	-1.06
2	9	-2.89	3.56	-10.29
5	4	0.11	-1.44	-0.16
8	3	3.11	-2.44	-7.59
7	3	2.11	-2.44	-5.15
12	2	7.11	-3.44	-24.46
2	7	-2.89	1.56	-4.51
4	7	-0.89	1.56	-1.39

$\Sigma = -64.57$



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** – equivalent to the covariance of the standardized variables

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.

Correlation and Causality

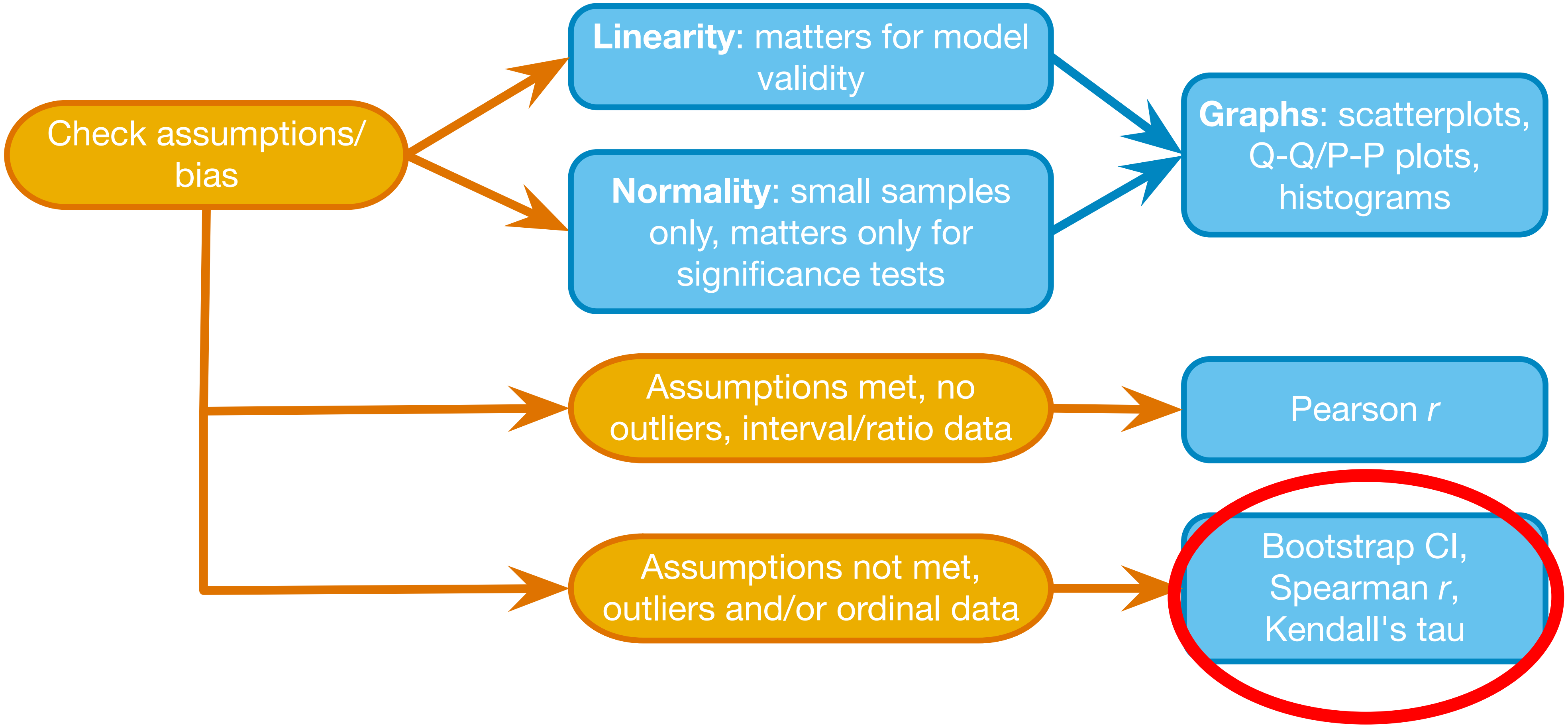
- The **third-variable problem**:

  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

Conducting Correlation Analysis



Nonparametric Correlation

- Spearman’s rho
  - Pearson’s correlation on the ranked data

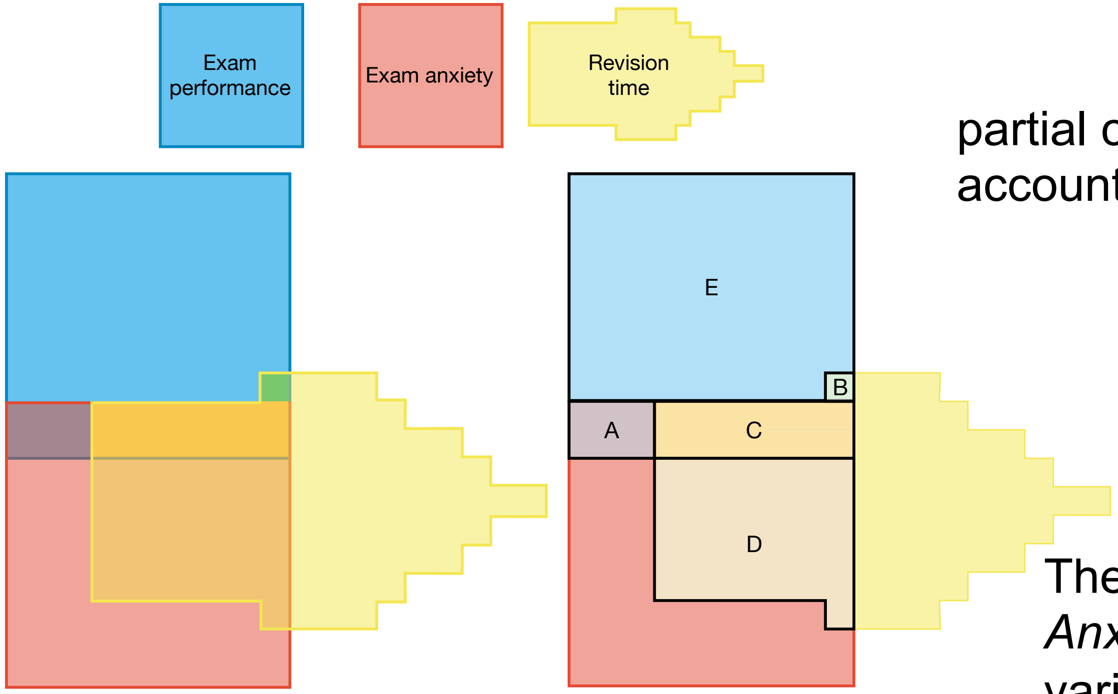
- Kendall’s tau
  - Better than Spearman’s for small samples

Partial Correlations

- Partial correlation:

  Measures the relationship between two variables, adjusting for the effect that a third variable has on them both

Partial Correlations



A = variance exam performance uniquely shared with exam anxiety (5.1%)

B = variance in exam performance uniquely shared with revision time (1.5%)

C = variance in exam performance shared by both exam anxiety and revision time (14.3%)

D = variance shared by exam anxiety and revision time but not exam performance (36%)

E = variance in exam performance not shared by any measured variable (79.1%)

A + C = variance shared by exam performance and exam anxiety (19.4%)

C + B = variance shared by exam performance and revision time (15.8%)

C + D = variance shared by revision time and exam anxiety (50.3%)

A + B + C = variance in exam performance accounted for by revision time and exam anxiety (20.9%)

partial correlation is the relationship between *X* and *Y* accounting for the overlap in *X* and *Z and Y* and *Z*

The **partial correlation** between *Performance* and *Anxiety* accounting for *Revision Time* is the unique variance in exam performance shared with exam anxiety (A) expressed as a proportion of the variance in exam performance not shared with revision time (A+E)

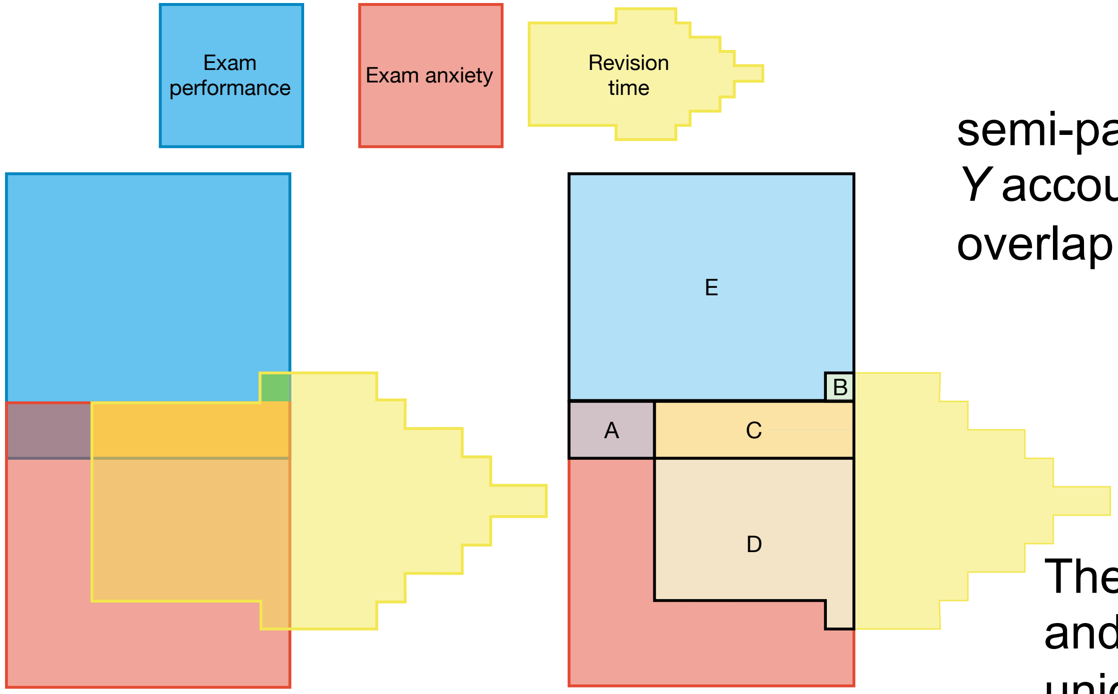
Semi-Partial Correlations

- Semi-partial correlation:

  A measure of the relationship between two variables while adjusting for the effect that one or more additional variables have on one of those variables.

  If we call our variables *x* and *y*, it gives us a measure of the variance in *y* that *x* alone shares

Semi-Partial Correlations



A = variance exam performance uniquely shared with exam anxiety (5.1%)

B = variance in exam performance uniquely shared with revision time (1.5%)

C = variance in exam performance shared by both exam anxiety and revision time (14.3%)

D = variance shared by exam anxiety and revision time but not exam performance (36%)

E = variance in exam performance not shared by any measured variable (79.1%)

A + C = variance shared by exam performance and exam anxiety (19.4%)

C + B = variance shared by exam performance and revision time (15.8%)

C + D = variance shared by revision time and exam anxiety (50.3%)

A + B + C = variance in exam performance accounted for by revision time and exam anxiety (20.9%)

semi-partial correlation is the relationship between *X* and *Y* accounting for the overlap in *X* and *Z*, but not the overlap in *Y* and *Z*

The **semi-partial correlation** between *Performance* and *Anxiety* accounting for *Revision Time* is the unique variance in exam performance (A) shared with exam anxiety expressed as a proportion of the variance in exam performance (A+C+E+B)

Categorical variables: Contingency Table

- Analyzing two or more categorical variables
  - The mean of a categorical variable is meaningless
    - The numeric values you attach to different categories are arbitrary
    - The mean of those numeric values will depend on how many members each category has.
  - Therefore, we analyze frequencies.
- An example
  - Can animals be trained to line-dance with different rewards?
  - Participants: 200 cats
  - Training
    - The animal was trained using either food or affection, not both)
  - Dance
    - The animal either learnt to line-dance or it did not.
  - Outcome:
    - The number of animals (frequency) that could dance or not in each reward condition.
  - We can tabulate these frequencies in a **contingency table**

A contingency table

TABLE 18.1 Contingency table showing how many cats will line-dance after being trained with different rewards

		Training		
		Food as Reward	***Affection as Reward***	***Total***
Could They Dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

Strength of the assocation between categorical variables can be quantified with a contingency coefficient or Cramer’s V – reviewed in today’s lab activity