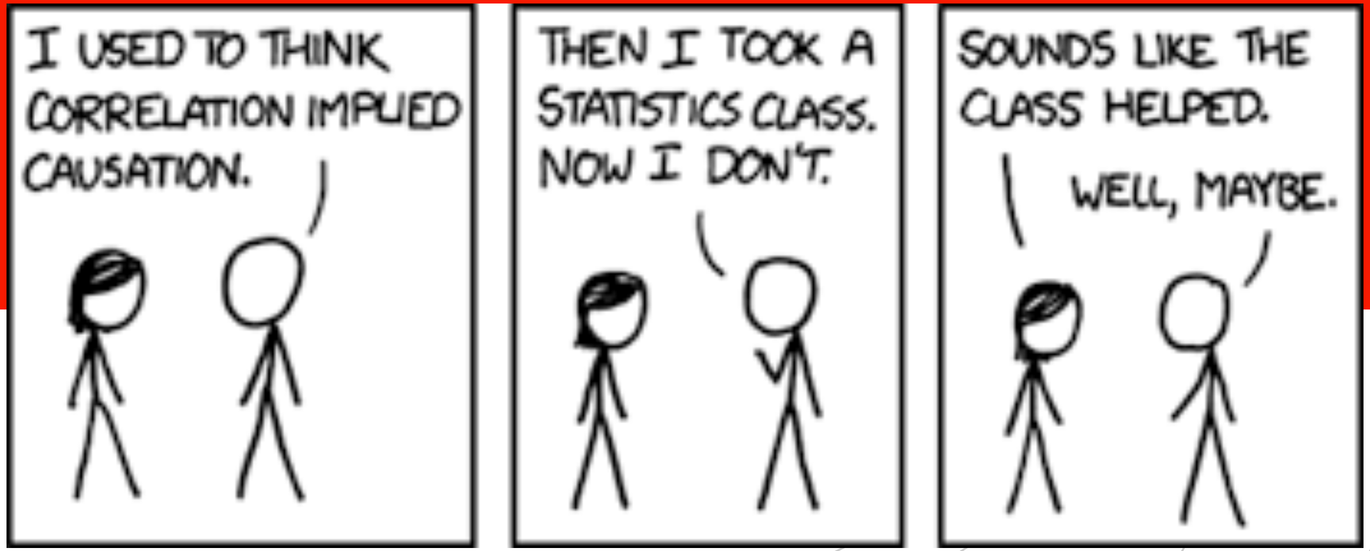
Analytical Issues With Correlated Data



Analytical Issues With Correlated Data

Outline

1. What is correlation?
2. How does it arise in our data?
3. Why is it important?

Analytical Issues With Correlated Data

Disclaimer: There will be a few formulas introduced in this lecture, but do not focus on them too much, we will be exploring them further in future lectures.

$Yi = \beta 0 + \beta 1 * Xi$

t-statistic: $\beta / \sqrt{[var(\beta)]}$

$Var (Y1 + Y2) = Var (Y1) + Var (Y2)$

1. What is correlation?

Fun with definitions!

1. What is correlation?

Poll: Collecting more data, (if budget allows) is always better:

- 1) Duh (Lawful good)
- 2) It depends (True Neutral)
- 3) You may be getting more than you bargained for (Chaotic Evil)

1. What is correlation?

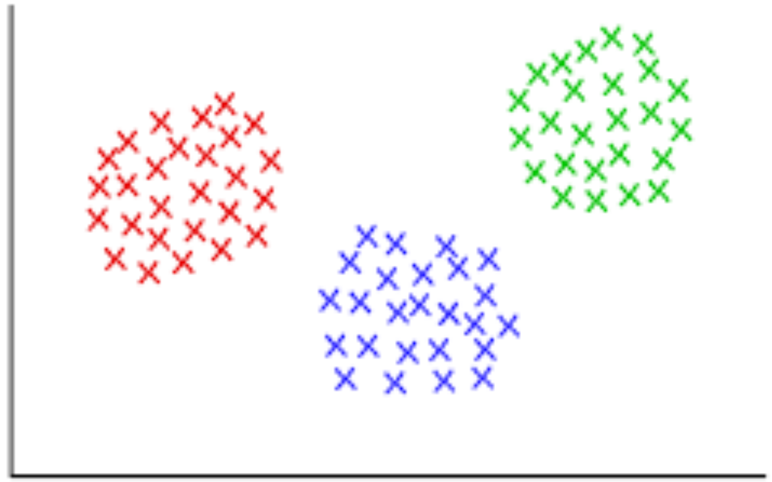
We can talk about correlation in 2 ways:

1) Relationship between two variables

1. What is correlation?

We can talk about correlation in 2 ways:

2) Correlation is the degree to which **outcomes** move together

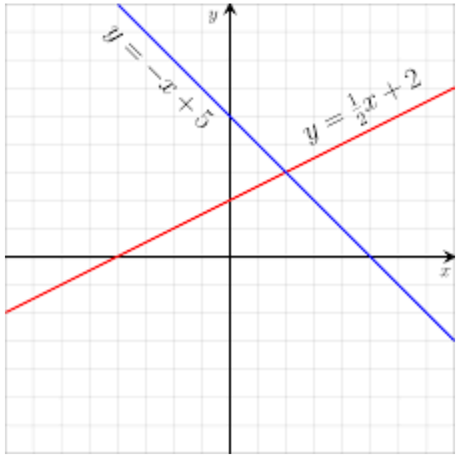


1. What is correlation?

- When we model data, we are often interested in determining the relationship between an outcome **Y** and one or more exposures **X**
- We often simplified this as the equation of a line

$$\mathbf{Yi = \beta 0 + \beta 1 * Xi}$$

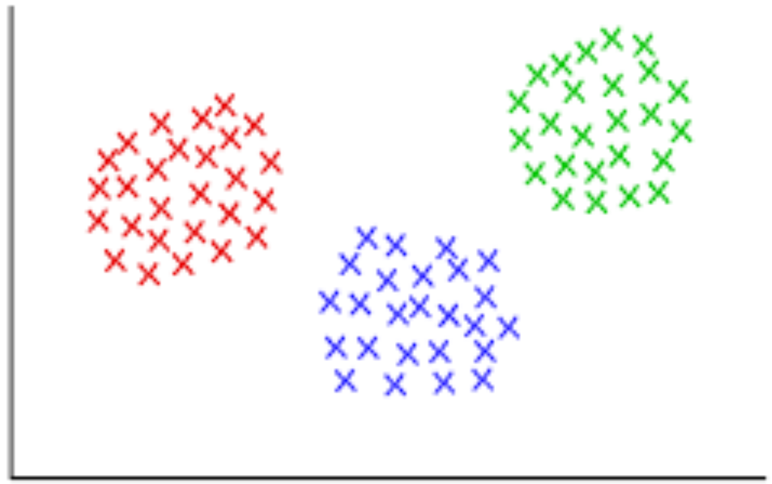
- Y is our outcome
- X is our exposure
- β0 is our intercept
- B1 is our slope



1. What is correlation?

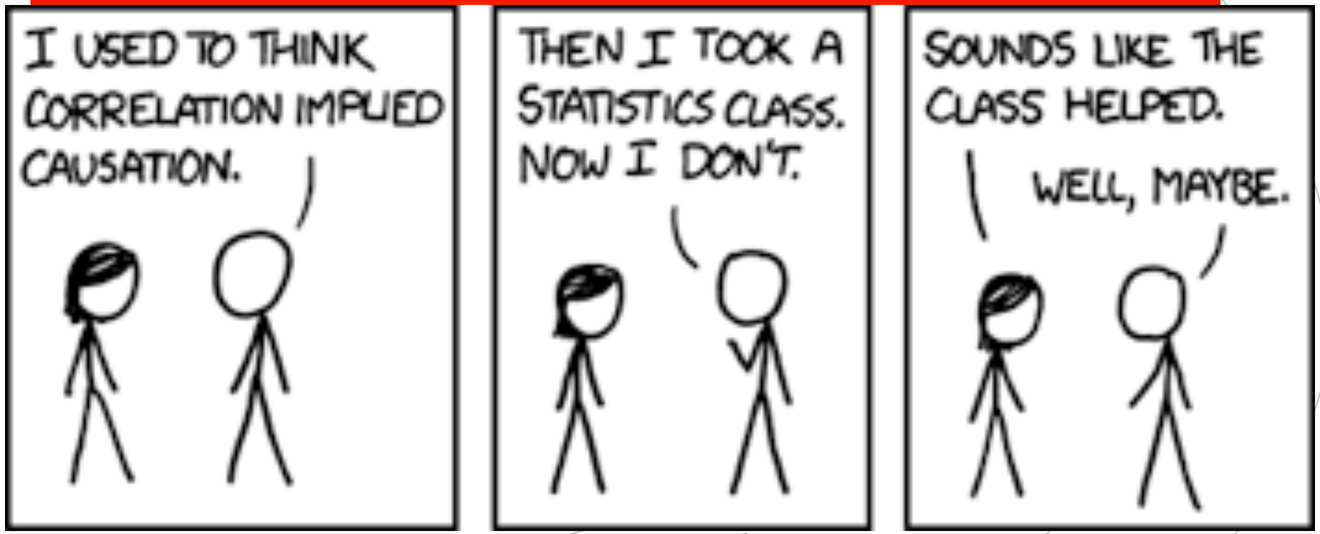
2) Correlation is the degree to which **<u>outcomes</u>** move together

$$\mathbf{Yi = \beta 0 + \beta 1 * Xi}$$



Each of these little dots is a Yi value

2. How does correlation arise in our data?

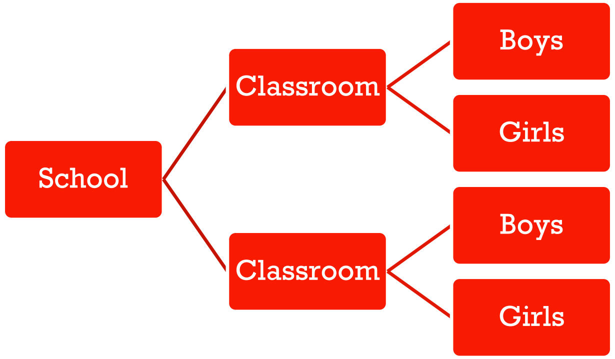


2. How does correlation arise in our data?

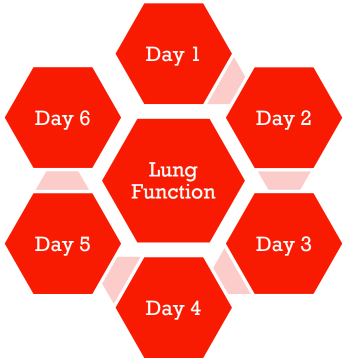
- We generally have two types of data: cross sectional and longitudinal.
  - Cross-sectional data is collected at a single timepoint. It is a "snapshot"
  - Longitudinal data is collected multiple times on a single entity over a period of time
- Example 1: If we collect the info on student IQ and GPA in a single at any one given time, it is cross-sectional
- Example 2: If we collect a student's GPA over time, it is longitudinal

2. How does correlation arise in our data?

HIERARCHICAL STRUCTURE



CLUSTERED STRUCTURE



2. How does correlation arise in our data?

- Some Examples
  - Looking at trends in grades over the semester
  - Comparing the number of goals scored by each striker within a league
  - Changes in biological samples assays between two sites

2. How does correlation arise in our data?

- Some Examples

  - Looking at trends in grades over the semester
    - Each student will have multiple grades (Yi) which would be correlated
  - Comparing the number of goals scored by each striker within a league
    - Each striker could score a similar amount of goals (Yi) during each game
  - Changes in biological samples assays between multiple sites
    - Samples from each site might be more similar to one another than compared to another site

Why is correlation so important in our data?

This is the part with formulas

3) Why is correlation so important in our data?

- Let's take another look at our old friend

$$\mathbf{Yi = \beta 0 + \beta 1 * Xi}$$

- You may recall some assumptions that came with it:
  - **L**: The relationship between X and Y is linear
  - **I**: Independence of the outcome values Y (or residuals)
  - **N**: All variables are normally distributed
  - **E**: Equality of residuals

3) Why is correlation so important in our data?

- Let's take another look at our old friend

$$\mathbf{Y_i = \beta_0 + \beta_1 * X_i}$$

- You may recall some assumptions that came with it:
  - L: The relationship between X and Y is linear
  - **I: Independence of the outcome values Y (or residuals)**
  - N: All variables are normally distributed
  - E: Equality of residuals

3) Why is correlation so important in our data?

- If two events Y1 and Y2 are independent, then
  - Probability (Y1 happens) & Probability (Y2 happens)

    = Probability (Y1 happens) * Probability (Y2 happens)
  - Covariance (Y1,Y2) = 0

- The variance of Y1 and Y2 is
  - Var (Y1 + Y2) = Var (Y1) + Var (Y2) + 2 * Covariance (Y1, Y2)
- If Y1 and Y2 are independent
  - Var (Y1 + Y2) = Var (Y1) + Var (Y2) + 0

3) Why is correlation so important in our data?

- We would like to go from this formula

$$\mathbf{Yi = \beta 0 + \beta 1 * Xi}$$

- To these formula (T-statistic and 95% CI of the mean)

T-statistic: β / $\sqrt{}$ [var(β)]

95% CI: β +/- 1.96* $\sqrt{}$ [var(β)]

- β is an estimate of our effect or mean value for Y in our sample. It can also be written as (**β = Y2-Y1**) if we are just looking at two observations. Looking at the t-statistic can tell us:
  - Is there an association?
  - What is the magnitude of the association?

3) Why is correlation so important in our data?

- Let's focus on the variance portion:

T-statistic = β / √ [var(β)]

T-statistic = Y2 – Y1 / √ [var(Y1,Y2)]

- If we assume Y1 and Y2 are independent

T-statistic = Y2 – Y1 / √ [var(Y1) + var(Y2)]

- However, if Y1 and Y2 are not independent

T statistic = Y2 – Y1 / √ [var(Y1) + var(Y2) + 2*covariance(Y1,Y2)]

3) Why is correlation so important in our data?

- If we use the naïve version of the T-statistic:

  T-statistic = Y2 – Y1 / √ [var(Y1) + var(Y2)]

- Instead of

  T-statistic = Y2 – Y1 / √ [var(Y1) + var(Y2) + 2*covariance(Y1,Y2)]

- Our interpretation of the results will change:
  - Is there an association? Biased, our T-statistic will be falsely increased!
  - What is the magnitude of the association? Biased, our T-statistic will be falsely increased!

3) Why is correlation so important in our data?

- Groups are not always constituted at random, but they can have some physical, geographic or social traits in common.
- We want to investigate correlation between 2 variables, and address correlation within one variable
- If we do not address correlation in our data, it can bias our analyses!