Statistics II

Week 2: **Causes & Effects**

Content for Today

1. Review of core concepts from lecture
2. Calculating NATE and ATE
3. Creating plots with `ggplot2`
4. Creating and plotting simulated data
5. Plotting DAGs with `ggdag`

But first - some note on the assignments

1. Please do not move the assignments in the cloned folder
   a. I.e., no sub-folders
2. When you're doing the peer reviews, the content for the header should:
   a. Use periods to denote decimals (not commas)
   b. Only contain a number (no brackets, etc.)
   c. Only use the original number of question bullets (if there are three, there should only be three in the assignment - please don't add more)

Like this:

```
---
title: "Reviewer Feedback Form, Assignment 1"
output: github_document
params:
  q1_score: 3
  q2_score: 3
  q3_score: 4.5
---
```

But first - some note on the assignments

3. Don’t forget that you need to clone the assignment solutions repo on GitHub, and view the HTML on your machine in order to see the full solutions.
4. What matters is the final answer your classmates get - not how they coded it.
    a. There are tons of ways to code in R - some are better than others, but not necessarily right or wrong.
    b. Though do feel free to suggest improvements in the feedback.
5. The output of the code must be visible in the HTML document! Without this, the submission is only worth half points (as described in the README).

Lecture Review

Potential Outcomes Framework

Key concept: Every individual in a sample could *theoretically* be exposed to two states with different potential outcomes. However, we can only ever observe one state. The other will always be a counterfactual, meaning we cannot observe causal effects at the individual level.

Therefore, we look at averages:

ATE: Expected mean value for those in the treatment group minus the expected mean value for those same individuals, had they not been treated.

NATE: The difference in the sample means on the observed outcome variable Y for the observed treatment and control groups.

Estimating Bias

We can define the ATE as a weighted average of the ATT and ATC:

ATT: The expected mean for those in the treatment group *with* treatment, minus the expected mean of those in the treatment group *without* treatment.

ATC: The expected mean for those in the control group *with* treatment, minus the expected mean of those in the control group *without* treatment.

$$ATE = \pi ATT + (1 - \pi)ATC$$

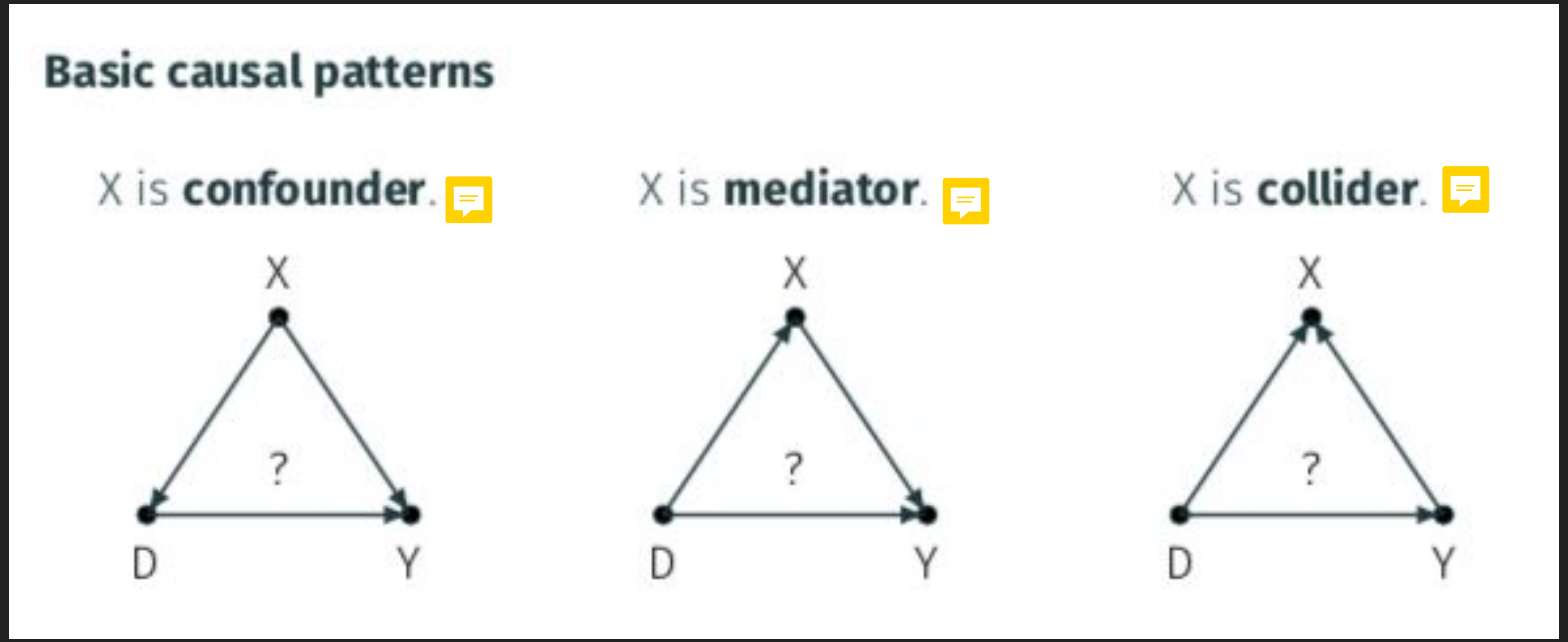
Where π is the proportion of subjects under treatment

Estimating Bias

Total bias of NATE as an estimate of ATE is simply NATE − ATE

It is made up of the **baseline bias** (difference in average outcome without treatment for the treatment and control groups) and **differential treatment effect bias (**the difference in the average treatment effect between the treatment and control groups, weighted by the proportion of the population in the control group).

Causal Graphs/DAGs



Randomized Experiments

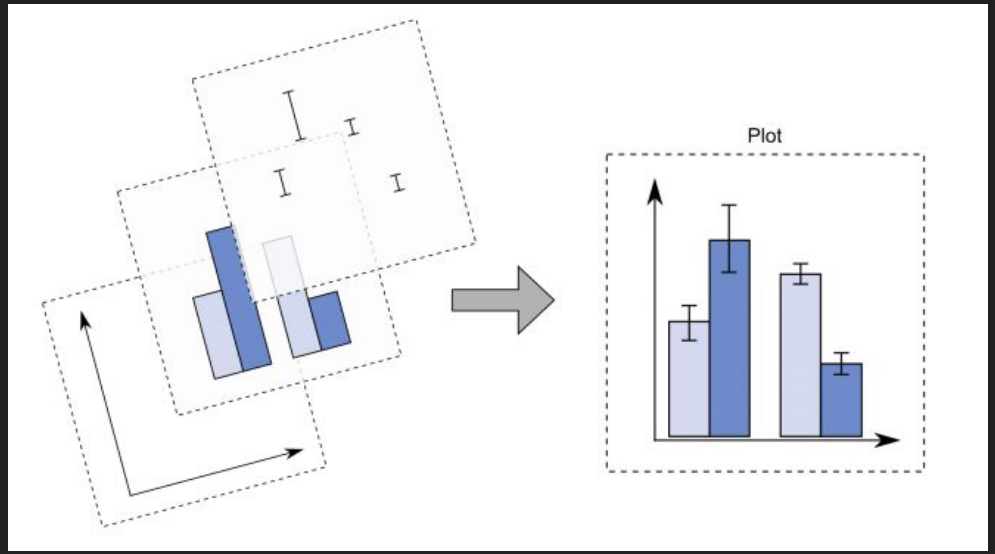
- In randomized experiments, assignment to the treatment or control groups *only depends on randomization*.
- We know that causal effects are identified if if potential outcomes are independent of the treatment assignment process (i.e. no self-selection bias - satisfaction of the independence/ignorability assumption).
    - Independence/ignorability assumption: The expected outcome of treatment for those treated would be the same for those in treatment or control group, and vice versa. *Basically, the fact that you get treatment or not does not affect your potential outcomes.*
- Therefore, if treatment is randomly assigned, the expected difference in Y between treatment and control is just the average of the above differences, and NATE = ATE.

Questions?

Plotting with `ggplot2`

ggplot2

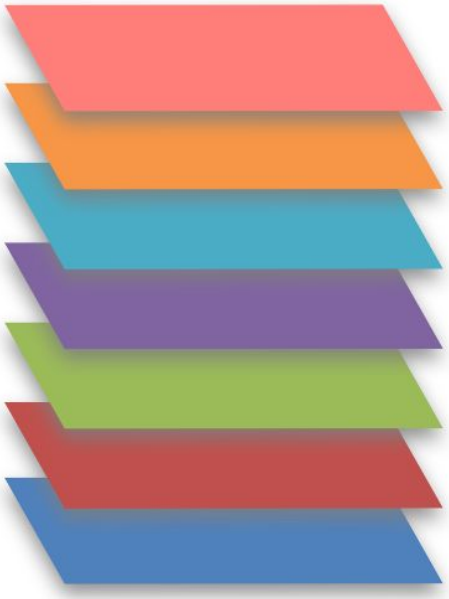
In `ggplot2`, a graph is made up of a series of layers



Download the cheat sheet: https://tinyurl.com/h5o9tfq

ggplot2

Describes all the non-data ink	Theme
Plotting space for the data	Coordinates
Statistical models & summaries	Statistics
Rows and columns of sub-plots	Facets
Shapes used to represent the data	Geometries
Scales onto which data is mapped	Aesthetics
The actual variables to be plotted	Data



ggplot2

Geometric objects are the visual elements such as bars and points: **`geom()`**

- There are many kinds of geoms, such as scatterplots (`geom_point`) or barplots (`geom_bar`)

The appearance and location of these geoms (such as size and color) are controlled by the **aesthetics properties**: `aes()`

- The variables you want to plot are are referred to here.

```
myGraph <- ggplot(data,
                  aes(x = variable_for_x_axis,
                      y = variable_for_y_axis)) +
                  geom()
```

ggplot2

Types of Geoms:

Scatterplot: `geom_point()`

Histogram: `geom_histogram()`

Barplot: `geom_bar()`

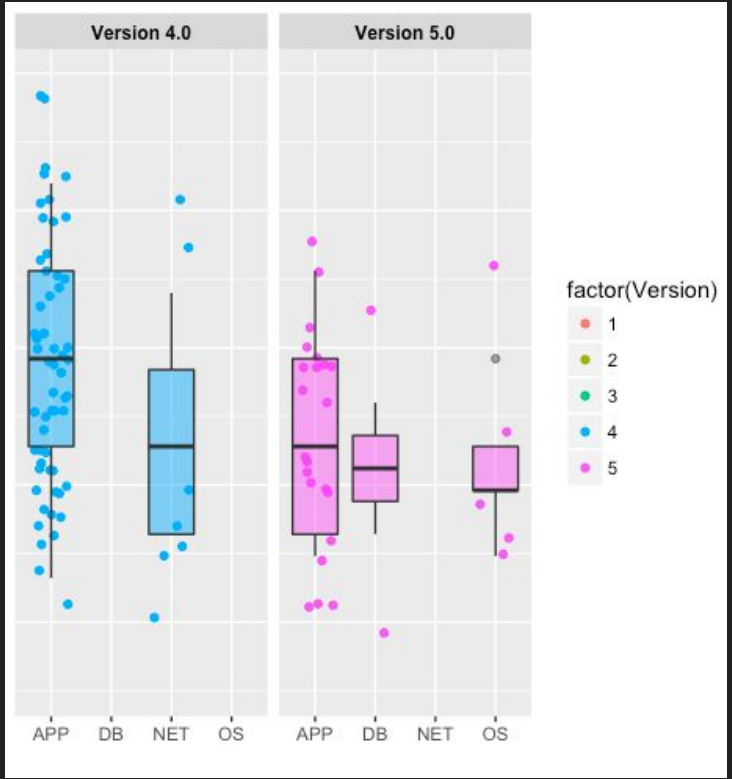
Boxplot: `geom_boxplot()`

Density: `geom_density()`

Adding a “linear regression” line:

`geom_smooth(model = lm)`

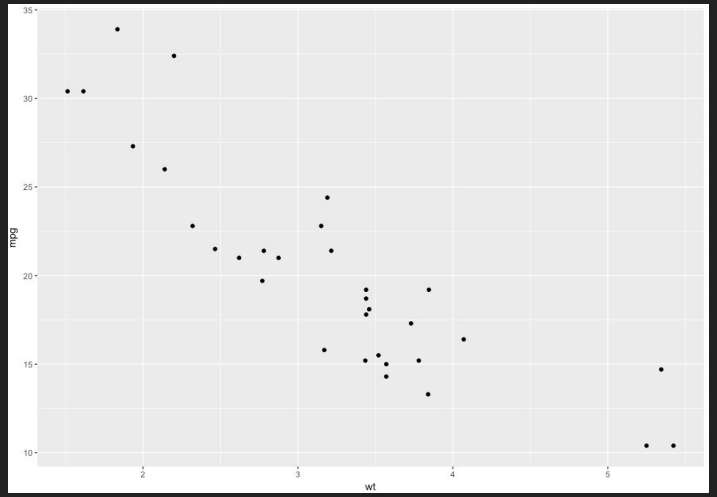
Or a combination of multiple.



ggplot2

Using the build-in `mtcars` data frame, we can plot a basic scatterplot:

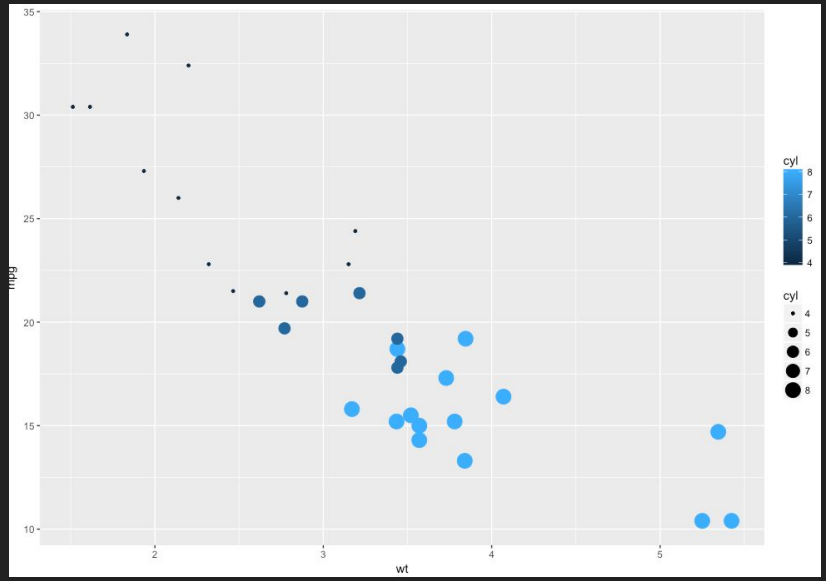
```
scatterplot <- ggplot(
    data = mtcars,
    aes(x = wt, y = mpg)) +
  geom_point()

print(scatterplot)
```



ggplot2

From here we can add extra design elements, like changing the color and size of the points based on cylinder size:

```
scatterplot <- ggplot(
    data = mtcars,
    aes(x = wt,
        y = mpg,
        col = cyl,
        size = cyl)) +
  geom_point()
```



ggplot2

There are lots of other things we can do, too!

Change the transparency using `alpha` and a number from 0-1: ex. `alpha = 0.6`

Add a theme: ex. `geom_bar() + theme_bw()`

Add labels and titles: ex. `+ xlab('X label')+ ylab('Y label') + ggtitle(' Title')`

Change the size and shape of lines, points, etc.

Zoom in on a certain part of a graph

And more :)