SCI 2025: Homework 3

Setup

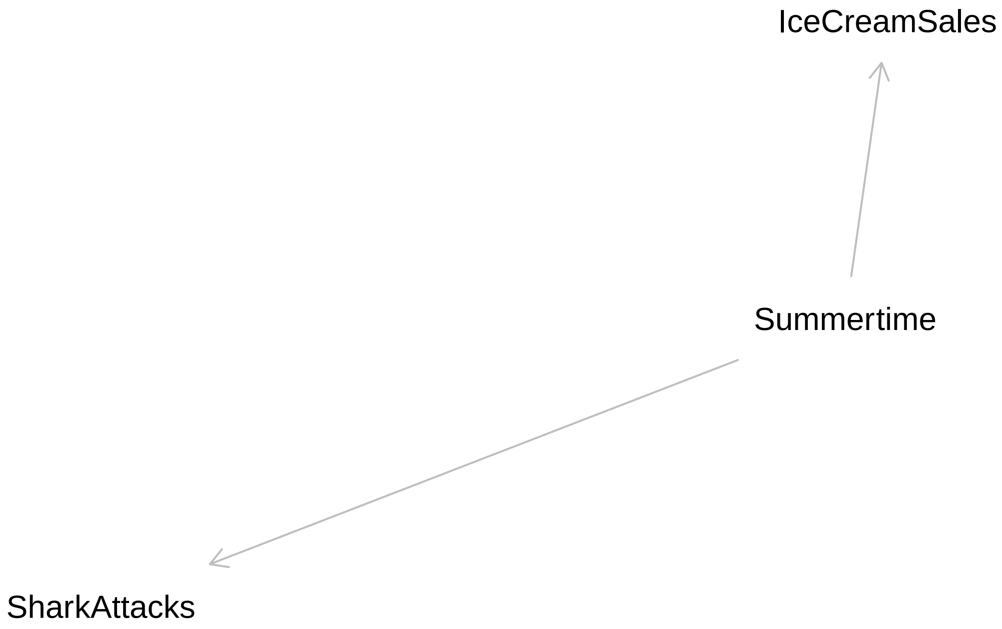
Simulation is one of the best ways to understand causal inference problems such as confounding. The reason is that, in any empirical dataset, we do not have access to the "ground truth" of the causal model or data generating process. By simulating data ourselves, we get to determine the truth (via parameter values) and thus have a benchmark for evaluating the performance of our models.

To demonstrate, I will simulate using the example of confounding from this week's lecture: shark attacks and ice cream sales. My causal assumptions are as follows:

- $Summertime \rightarrow SharkAttacks$
- $Summertime \rightarrow IceCreamSales$

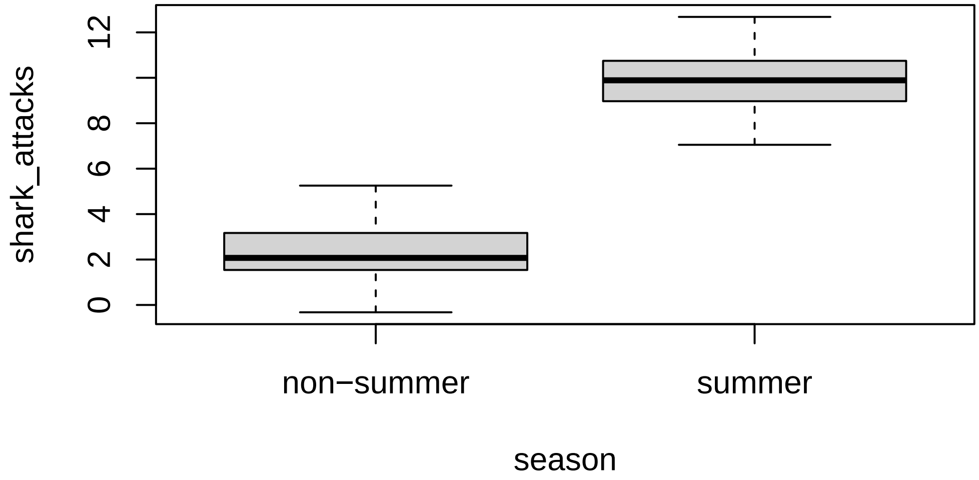
The full DAG looks like this:

```r
library(dagitty)
DAG <- dagitty("dag {
  Summertime -> SharkAttacks
  Summertime -> IceCreamSales
}")
plot(DAG)
```

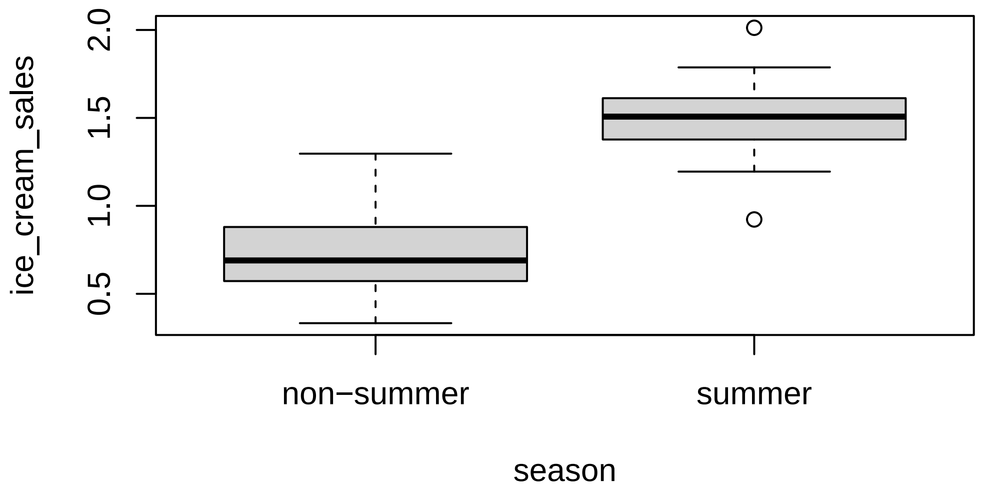


Let's imagine that we are dealing with national-level data from Australia. We have 30 years of data on the number of shark attacks and ice cream sales in the summer and non-summer months.

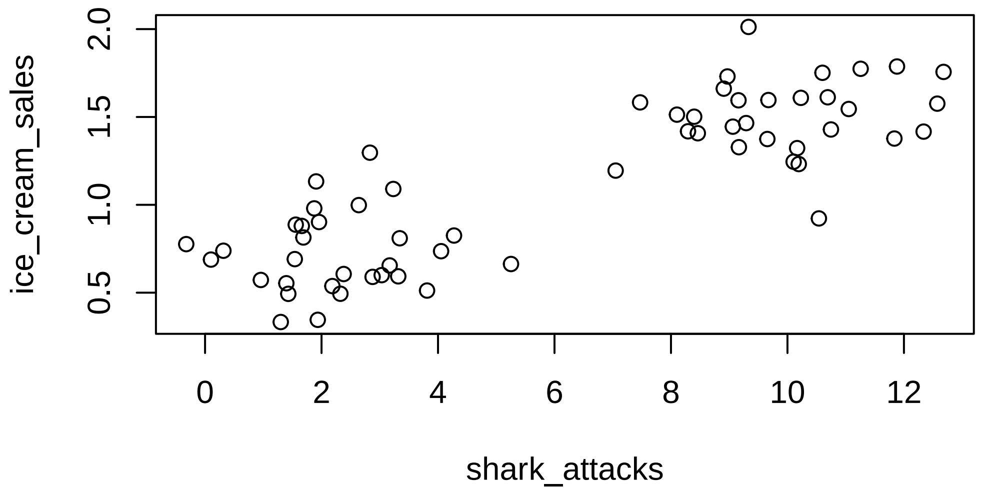
```
# Set the seed for reproducibility
set.seed(123)

n_years <- 30

season <- rep(c("summer", "non-summer"), each = n_years)

num_attacks_summer <- 10 # average number of shark attacks in the summer

num_attacks_non_summer <- 2 # average number of shark attacks in the non-summer

# simulate the number of shark attacks, conditional on the season
shark_attacks <- rnorm(length(season),
 ifelse(season == "summer", num_attacks_summer, num_attacks_non_summer),
  1.5)

boxplot(shark_attacks ~ season)
```



```
# simulate the number of ice cream sales, conditional on the season
# average number of ice cream sales in the summer, billions
avg_sales_summer <- 1.5
# average number of ice cream sales in the non-summer, billions
avg_sales_non_summer <- 0.75

# simulate the number of ice cream sales, conditional on the season
ice_cream_sales <- rnorm(length(season),
 ifelse(season == "summer", avg_sales_summer, avg_sales_non_summer),
  0.25)

boxplot(ice_cream_sales ~ season)
```



```
# Now, see the relationship between ice cream sales and shark attacks
plot(ice_cream_sales ~ shark_attacks)
```



Question 1

Come up with an example from your own field of research where you might expect to see confounding (i.e., a backdoor path between the predictor and outcome). Do not use examples from the lecture, previous homeworks, or the textbook. Describe your causal assumptions (not statistical assumptions), either verbally or via a DAG.

Question 2

Based on the causal model you described in Question 1, simulate data from the model. This will require you moving from qualitative causal assumptions to quantitative statistical assumptions. You may justify your choices of parameter values based on your knowledge of the real world, or choose arbitrary/standardized values. Most important is that your choice of parameter values is consistent with the causal assumptions you made in Question 1. Inspect the data you have simulated and make some plots to visualize the relationships between the variables.

Question 3

Fit a regression model to the data you simulated in Question 2. However, *this model should omit any confouning variables.* Print the model summary, and compare the model estimates to the true parameter values you used to simulate the data.

Question 4

Now, fit a new regression model to the data you simulated in Question 2. However, *this model should include the confouning variable.* Print the model summary, and compare the model estimates to the true parameter values you used to simulate the data.

Question 5

Re-run the simulations and model fitting steps from Questions 2-4, but modify the effect of the confouning variable. Describe the effect of this change on the model estimates.

Question 6

Re-run the simulations and model fitting steps from Questions 2-4, using *the same set of parameter values that you used the first time.* But this time, set a different random seed number (see my example of `set.seed()` in the setup). Describe the effect of this change on the model estimates. Can you explain why the estimates did/or did not change?