Causal Inference and Missing Data
Computer exercises week 1

Exercise 1 (Different research questions)

Suppose you have access to individual patient data from electronic health records of all patients who were admitted to the intensive care unit (ICU) of the LUMC hospital between 2010 and 2020. Data contains info on admission date, discharge date, discharge status (alive / deceased), age at admission, underlying disease, disease severity, use of mechanical ventilation at ICU yes/no and use of platelet transfusion at ICU yes/no.

Formulate three hypothetical research questions that one could study with such data: one descriptive question, one predictive question and one causal question.

Exercise 2 (Potential outcome simulation, Part 3.2.1 r-causal.org)

Let's suppose some happiness index, from 1-10 exists. We are interested in assessing whether eating chocolate ice cream versus vanilla will increase happiness. We have 10 individuals with two potential outcomes for each, one is what their happiness would be if they ate chocolate ice cream, (defined as y_chocolate in the code below), and one is what their happiness would be if they ate vanilla ice cream (defined as y_vanilla in the code below). We can define the true causal effect of eating chocolate ice cream (versus vanilla) on happiness for each individual as the difference between the two. Generate such data using the following code.

```
data <- data.frame(
  id = 1:10,
  y_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
)
```

2a: Calculate the individual causal effect for each person

2b: Introduce notation for the potential outcomes and for the exposure

2c: Formulate the average treatment effect (expressed as a difference) using the potential outcome notation and in words

2d: Calculate the mean potential outcome in each group and the *causal* average treatment effect

Exercise 3 (Observational exposure, Part 3.2.1 r-causal.org)

In reality, we cannot observe both potential outcomes, in any moment in time, each individual in our study can only eat one flavor of ice cream. Suppose we let our participants choose which ice cream they wanted to eat and each chose their favorite (i.e. they knew which would make them "happier" and picked that one.

3a. Generate the observed exposure (chocolate if the participant prefers chocolate ice cream and 'vanilla' if the participant preference vanilla ice cream. Then generate a new variable with the observed outcome for each participant. Use the potential outcome data of Exercise 2.

3b: Calculate the *observed* average treatment effect under this design. Is it the same as the causal average treatment effect calculated in Exercise 2?

3c: Which of the causal assumptions discussed in the lecture was violated in the data analysis performed in question 3b?

3d: Is there a way to perform a corrected analysis on these data that circumvents the problems introduced by the impact of the participants' preferences?

Exercise 4 (Randomised exposure, Part 3.2.1 r-causal.org)

Now suppose we randomly allocated participants to one of the two ice cream flavours. You may use the code below to mimic such an experiment.

```
## we are doing something *random* so let's set a seed so we always observe the same result each time we run the code
set.seed(11)
data_observed <- data |>
 mutate(
  # change the exposure to randomized, generate from a binomial distribution with a probability 0.5 for being in either group
  exposure = case_when(
   rbinom(10, 1, 0.5) == 1 ~ "chocolate",
   TRUE ~ "vanilla"
  ),
  observed_outcome = case_when(
   exposure == "chocolate" ~ y_chocolate,
   exposure == "vanilla" ~ y_vanilla
  )
 ) |>
 # we can only observe the exposure and one potential outcome
 select(id, exposure, observed_outcome)
```

4a: What is the *observed* treatment effect from this from the randomized experiment?

4b: Compare the *observed* treatment effect to the *causal* average treatment effect calculated in Exercise 2, explain any differences.

4c: Are the three causal assumptions met with this randomized procedure?

Exercise 5 (Causal assumptions, Part 3.3.1 r-causal.org)

We continue with the randomized design, but now suppose that there were in fact two containers of chocolate ice cream, one of which was spoiled. Therefore, having an exposure “chocolate” could mean different things depending on where the individual’s scoop came from (regular chocolate ice cream, or spoiled chocolate ice cream). You can use the code below.

```
data <- data.frame(
  id = 1:10,
  y_spoiledchocolate = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
  y_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
) |>
  mutate(causal_effect = y_chocolate - y_vanilla)

set.seed(11)
data_observed <- data |>
  mutate(
    exposure_unobserved = case_when(
      rbinom(10, 1, 0.25) == 1 ~ "chocolate (spoiled)",
      rbinom(10, 1, 0.25) == 1 ~ "chocolate",
      TRUE ~ "vanilla"
    ),
    observed_outcome = case_when(
      exposure_unobserved == "chocolate (spoiled)" ~ y_spoiledchocolate,
      exposure_unobserved == "chocolate" ~ y_chocolate,
      exposure_unobserved == "vanilla" ~ y_vanilla
    ),
    exposure = case_when(
      exposure_unobserved %in% c("chocolate (spoiled)", "chocolate") ~ "chocolate",
      exposure_unobserved == "vanilla" ~ "vanilla"
    )
  ) |>
  select(id, exposure, observed_outcome)
```

5a. What is the observed treatment effect now?

5b. Which of the causal assumptions discussed in the lecture is violated here?

Exercise 6

Now suppose each individual has a partner, and their potential outcome depends on both what flavor of ice cream they ate and what flavor their partner ate. For example, in the simulation below, having a partner that received a different flavor of ice cream increases the happiness by two units. See code below

```
data <- data.frame(
 id = 1:10,
 partner_id = c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5),
 y_chocolate_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
 y_chocolate_vanilla = c(6, 6, 8, 7, 8, 7, 8, 9, 7, 8),
 y_vanilla_chocolate = c(3, 5, 6, 7, 7, 8, 10, 8, 5, 7),
 y_vanilla_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
)

set.seed(11)
data_observed <- data |>
 mutate(
  exposure = case_when(
   rbinom(10, 1, 0.5) == 1 ~ "chocolate",
   TRUE ~ "vanilla"
  ),
  exposure_partner =
   c("vanilla", "vanilla", "vanilla", "chocolate", "chocolate", "vanilla", "vanilla", "vanilla", "vanilla",
"chocolate"),
  observed_outcome = case_when(
   exposure == "chocolate" & exposure_partner == "chocolate" ~ y_chocolate_chocolate,
   exposure == "chocolate" & exposure_partner == "vanilla" ~ y_chocolate_vanilla,
   exposure == "vanilla" & exposure_partner == "chocolate" ~ y_vanilla_chocolate,
   exposure == "vanilla" & exposure_partner == "vanilla" ~ y_vanilla_vanilla
  )
 ) |>
 select(id, exposure, observed_outcome)
```

6a: What is the observed treatment effect now?

6b. Which of the causal assumptions discussed in the lecture is violated here?

Exercise 7

This is the first exercise of the group assignment.

- Form a group, keeping in mind that different skills are needed: translating questions from the applied world to statistical formulas, mathematical skills, programming skills and writing skills.
- Register with your group on Brightspace
- Perform the first (week 1) part of the assignment