Causal Inference I
Exercises week 2

Exercise 1 . Design and protocol

Below you find part of an abstract of a paper where the effect of a psychological artificial intelligence chatbot, named Tess, on depression and well-being of students was studied (Fulmer et al JMIR Ment Health. 2018 Dec 13;5(4):e64. doi: 10.2196/mental.9782.)

Objective: This study aimed to assess the feasibility and efficacy of using an integrative psychological AI, Tess, to reduce self-identified symptoms of depression and anxiety in college students.

Methods: In this randomized controlled trial, 75 participants were recruited from 15 universities across the United States. All participants completed Web-based surveys, including the Patient Health Questionnaire (PHQ-9), Generalized Anxiety Disorder Scale (GAD-7), and Positive and Negative Affect Scale (PANAS) at baseline and 2 to 4 weeks later (T2). The 2 test groups consisted of 50 participants in total and were randomized to receive unlimited access to Tess for either 2 weeks (n=24) or 4 weeks (n=26). The information-only control group participants (n=24) received an electronic link to the National Institute of Mental Health's (NIMH) eBook on depression among college students and were only granted access to Tess after completion of the study.

a. What kind of study design best describes this study?
b. Can you formulate the first 6 key items of a protocol for this study?
c. Do you think that the exposed and unexposed group are exchangeable? Why (not)?
d. Suppose that some participants who were assigned to have unlimited access to Tess, never used Tess during the four weeks of the study. Discuss how you would handle the outcomes of these participants in the analysis.

Exercise 2 Design and protocol

3. Based on Jewell Statistics for Epidemiology , Chapman and Hall 2007

Helms et al (2003) investigate the role of serious gastrointestinal infections on 1-year mortality by using Danish registry systems. The investigators sampled all cases of culture confirmed infections listed with the Danish registry of enteric pathogens. For each case another 10 randomly selected individuals were chosen from the Danish civil registration system that assigns a personal identification number to all citizens of Denmark. These control individuals had to be alive on the date the infection sample was received for the case. The registration system was then used to obtain vital status and date of any death within 1 year from infection date in both cases and controls

a. What kind of study design best describes this study?
b. Can you formulate the first 6 key items of a protocol for a target trial for this study?
c. Do you think that the exposed and unexposed students are exchangeable? Why (not)?

Exercise 3 Confounding (Part 4.3.1 r-causal.org)

3. In this exercise we will explore how the effect of a confounder, a variable which is common cause of exposure and outcome, distorts the relation between exposure and outcome, and we will explore a simple method to handle confounding.

a. Use R to simulate a binary confounder variable (with levels 0 and 1) with n=10,000 observations, using a binomial distribution with a probability 0.5 for having either level.

b. Generate the exposure variable from a binomial distribution with the probability of exposure dependent on the value of the confounder. The probability of exposure = 1 is 0.75 if the confounder is1, and 0.25 if the confounder is 0.

c. Generate potential outcome data with an "true" average treatment effect of 0. The potential outcome if exposure =0 depends on the value of the confounder but not on the value of the exposure. Add a random error term with mean 0 and standard deviation 1 as follows:

```
y0 <- confounder + rnorm(n)
```

Calculate also y1, the potential outcome if exposure is set to 1. Because the true effect is 0, y1 is identical to y0.

d. In practice we only observe one of the potential outcomes. Generate the observed outcome, which equals y1 if exposure = 1 and y0 if exposure = 0.

e. Make a cross table for the confounder versus the exposure. Are the exposure groups exchangeable?

f. Calculate the observed difference in mean outcome between the two exposure groups. Can we interpret this value causally?

g. One way to address confounding is by estimating the exposure effect separately within each confounder level. Estimate the difference in mean outcome separately for those with confounder = 1 and 0. Then average these two effect estimates. Can we interpret this average difference causally?

The method we used in g. to solve the fact that the two groups are not exchangeable is known as stratification. We are stratifying by the confounder(s) and estimating the causal effect within each stratum. To get an overall average causal effect we are averaging across the strata. This can be a great tool if there are very few confounders.

Exercise 4 Selection bias

4. In this exercise we will explore how selection of a non-representative sample from the population can introduce bias.

Consider a population where the relation between diabetes as exposure and cancer as outcome is studied. The following data are observed:

	Cancer (Y=1)	No cancer(Y=0)	Total
Diabetes (X=1)	10	90	100
No Diabetes (X=0)	1000	9000	10 000
Total	1010	9090	10 100

a. We assume that the exposed and non-exposed group are exchangeable. Estimate the risk difference, the risk ratio and the odds ratio.

b. Both having cancer and having diabetes increase the risk of being admitted to a hospital. Assume that in this population 1 % of the people with no diabetes and no cancer is in the hospital; 10% of the people with only diabetes; 10% of the people with only cancer and 20% of the people with both cancer and diabetes. Fill in the numbers of the cross-table for the subgroup of people who are in hospital:

In hospital:

	Cancer (Y=1)	No cancer(Y=0)	Total
Diabetes (X=1)			
No Diabetes (X=0)			
Total			

c. Calculate the risk difference, risk ratio and odds ratio in the subgroup of people who are in hospital. What do you observe?

Exercise 5

Make the second (week 2) part of the group assignment