Causal Inference I
Answers exercises week 2

Exercise 1 . Design and protocol

a. What kind of study design best describes this study?

This is a randomized controlled trial.

b. Can you formulate the first 6 key items of a protocol for this study?

Eligibility criteria: this is not clearly stated in the abstract. The participants are students with depression or anxiety complain

Exposure definition: There are 3 exposure groups. 1: receive unlimited access to Tess for 2 weeks, 2: receive unlimited access to Tess for 4 weeks and 0: control: an electronic link to the National Institute of Mental Health's (NIMH) eBook on depression.

Assignment procedures: random assignment

Follow-up period: 4 weeks

Outcome definition: There are three different outcome scores: the Patient Health Questionnaire (PHQ-9), Generalized Anxiety Disorder Scale (GAD-7), and Positive and Negative Affect Scale (PANAS)

Causal contrast of interest: There are different contrast of interest. For example, let Y be the PHQ-9 score, then a causal contast of interest could be : E(Y(1))- E(Y(0)). This is the difference between the average PHQ-9 score at 4 weeks had all students receive unlimited access to Tess for 2 weeks, and the average PHQ-9, had they not accessed Tess at all?

c. Do you think that the exposed and unexposed group are exchangeable? Why (not)?

The groups are randomize which will result in exchangeable groups. However the assignment of the exposure is not blinded. There may be a placebo effect.

d. Suppose that some participants who were assigned to have unlimited access to Tess, never used Tess during the study. Discuss how you would handle the outcomes of these participants in the analysis.

In an intention to treat analysis, all observations are analysed in the group they were originally be assigned to. This analysis considers the effect of offering Tess to students with depressive complains.

Exercise 2 Design and protocol

a. What kind of study design best describes this study?

This is an observational cohort study . For every person with an infection (X=1), 10 persons without an infection (X=0) are followed for a year.

a. In this study the binary exposure is serious gastrointestinal infections (defined as culture confirmed infections with nontyphoidal Salmonella, Campylobacter, Yersinia enterocolitica, or Shigella). The binary outcome is one year mortality.

b. People who have serious gastrointestinal infections, may differ from people who do not have these infections (for example smoking behavior, alcohol use, may increase the probability to obtain these infections). Therefore these groups are probably not exchangeable.

c. A case-control study. Collect data from a sample of people who died (Y=1), and who did not(Y=0) **(**maybe match cases and controls on age), and look back in time whether they had any severe gastrointestinal infections. Question to you: which design do you prefer, the cohort study, or the case-control study?

b. Can you formulate the first 6 key items of a protocol for a target trial

Eligibility criteria: the total Danish population

Exposure definition: In this study the binary exposure is serious gastrointestinal infections (yes/no).

Assignment procedures: there is no randomization here

Follow-up period: all people are followed for one year

Outcome definition: Death is the outcome of interest

Causal contrast of interest: A causal contrast of interest could be E(Y(1))- E(Y(0)). As Y is binary, this is a risk difference (because E(Y(1)) = the probability that Y(1)). This is the difference between the risk of death within a year had all Danish people encountered an infection versus the risk of death if no one was infected.

Another estimand could be the average effect in those being infected : E(Y(1)|X=1)- E(Y(0)|X=1). This indicates the excess risk of death by infection in those who were infected. We will discuss different estimands later in this course.

c. Do you think that the exposed and unexposed group are exchangeable? Why (not)?

Exercise 3 Confounding

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.

```
set.seed(12345)

n <- 10000
confounder <- rbinom(n, 1, 0.5)
```

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.

```
p_exposure <- ifelse(confounder ==1, 0.75, 0.25)
exposure <- rbinom(n, 1, p_exposure)
```

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.

```
Y1 <- y0
```

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

```
outcome <- ifelse(exposure ==1, y1,y0)
```

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

```
table (confounder, exposure)
```

There is no exchangeability, the percentage of observations with confounder =1 is much larger in the exposed group

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

```
mean(outcome[exposure==1])- mean(outcome[exposure==0])
```

The calculated difference is around 0.5. This is not a causal difference because there is no exchangeability

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?

```
# mean difference in those with confounder=1
mean(outcome[exposure==1 & confounder==1])- mean(outcome[exposure==0&
confounder==1])

# mean difference in those with confounder=0
mean(outcome[exposure==1 & confounder==0])- mean(outcome[exposure==0&
confounder==0])
```

All observations With the same value of the confounder, have the same probability to receive the exposure. Therefore there is exchangeability within the confounder classes. Therefore the average difference calculated here can be causally interpreted.

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 causal risk difference, the risk ratio and the odds ratio.

The Risk difference =0, the risk ratio =1, the odds ratio is 1

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 crosstable for the subgroup of people who are in hospital:

In hospital:

	Cancer (Y=1)	No cancer(Y=0)	Total
Diabetes (X=1)	2	9	11
No Diabetes (X=0)	100	90	190
Total	102	99	201

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

The RD= 2/11 -100/190= -0.34

The RR = (2/11)/ (100/190)

OR = (2/9)/(100/90) = 0.20

In the hospital a negative association between diabetes and cancer is observed. This spurious association is caused by selection bias. This will be discussed in the next class

Exercise 5

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