Causal Inference - Assignment Part2

Emilia Löscher 8470014 Kyuri Park 5439043

12 April, 2022

1. Data Description & Pre-analysis

We are given the "*glorious_data*" from Maaike, Hsuan and Daniel.
The dataset contains 5 variables:

- `CBD` : Cannabidiol Consumption in ml/day (range from 0 - 120, mean = 55)
- `EtD` : Eating Disorder Scale (scale from -30 - 30, mean = 0)
- `ChP` : Chronic Pain (scale from 0 - 100, mean = 50)
- `LfS` : Life Satisfaction (scale from 0-100, mean = 50)
- `Anx` : Anxiety (scale from 0 - 120, mean = 60)

Based on the following evidence we have found, we believe that the DAG in *Figure1* below is the most realistic and plausible.

- `CBD` → `ChP`: Hill (2015) found that use of cannabinoids is effective in reducing chronic pain and neuropathic pain.
- `CBD` → `ChP` → `LfS`: Aggarwal et al. (2013) found that the self-rated quality of life with the chronically ill patients who used cannabis is higher than that of the chronically ill patients who did not use cannabis.
- `CBD` → `Anx`: According to Blessing et al. (2015), the current evidence strongly supports CBD as a treatment for multiple anxiety disorders.
- `CBD` → `Anx` → `EtD`: Alzaher (2022) suggests that CBD has efficacy in reducing stress and promoting relaxation. Accordingly, by reducing the accompanying symptoms of eating disorders such as anxiety and depression, CBD helps with eating disorder recovery.
- `EtD` → `LfS`: According to Claden et al. (2020), individuals who have/had eating disorders scored lower on the Satisfaction with Life Scale (SWLS) than the general population.
- `Anx` → `LfS`: Serin et al. (2010) showed that the students' level of life satisfaction is significantly predicted by their anxiety and socio-economic level (The less anxiety, the higher the life satisfaction).

Note: All the code for the figures/analysis can be found in the *Appendix* at the end of the document.

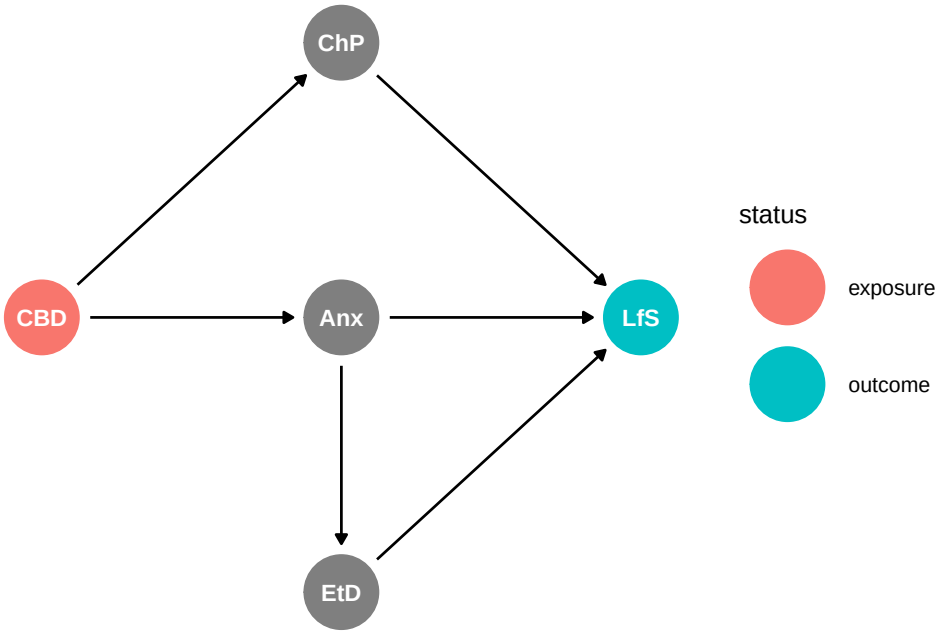

Figure 1: Our Expectation: DAG structure

2. Use the PC-algorithm on the data your received to 'discover' the structure of the causal system.

2a. Provide the CP-DAG. Discuss the main findings.

Figure2 shows the CP-DAG estimated with PC-algorithm. It is quite different from what we expected (*Figure1*). First of all, the `Anx` (anxiety) turned out to be a collider, which is surprising. We think that the relationship between `Anx` (anxiety) and `LfS` (life satisfaction) would be `Anx` (anxiety) $\rightarrow$ `LfS` (life satisfaction), rather than `Anx` $\leftarrow$ `LfS`, since anxiety is more likely to be a factor that affects life satisfaction, not the other way around. In addition, as we mentioned above in *Q1*, anxiety is found to be a significant predictor of life satisfaction (Serin et al., 2010). Hence, the edge: `Anx` $\leftarrow$ `LfS` is really in contrast to our expectation. `CBD` $\rightarrow$ `Anx` is detected as we expected, which is also supported by a previous study (Blessing et al.,2015).

Secondly, `EtD` (eating disorder) is found to have no relationship with any other variables. We suspect that this may have something to do with the fact that eating disorder is measured on a very different scale (-30 to 30) compared to the others (e.g., 0 to 100, 0 to 120). Also, we think that it makes sense substantively that the relationship between eating disorder and the other variables is relatively weaker, as it seems to be the least relevant variable in this DAG. Therefore, the absence of edges between eating disorder and the other variables is in a way sensible.

Lastly, the edges between `CBD` – `ChP` – `LfS` are not oriented, which accordingly results in three different possible DAGS. As seen in *Figure3*, there are three Markov-equivalent DAGs, which specify different relationships between `CBD` – `ChP` – `LfS`. `CBD` $\rightarrow$ `ChP` $\leftarrow$ `LfS` is ruled out by PC algorithm, because in that case `ChP` would be another collider.

2b. Select and provide the DAG you think is best from the equivalence set, that is, that you feel is most realistic/sensible given the information you received on the variables/overall context. Why do you prefer this DAG? Are there things about it that you find unrealistic?

Given what we expected (see *Figure1*), we think the DAG in ***Figure3 (c)*** makes the most sense: `ChP` (chronic pain) as a mediator between `CBD` and `LfS`. First, intuitively, `CBD` causing `ChP` (e.g., cannabidiol consumption reduces the chronic pain) and correspondingly `ChP` causing `LfS` (e.g., lower chronic pain increases the life satisfaction) seems to be the most sensible/plausible relationship. Additionally, Aggarwal et al. (2013) found that the self-rated life quality is higher with the chronically ill patients who used cannabis than that of who did not, which supports `CBD` $\rightarrow$ `ChP` $\rightarrow$ `LfS`.

The DAG in *Figure3 (a)* is excluded, as it does not make sense that the life satisfaction (`LfS`) causes the chronic pain (`ChP`). The DAG in *Figure3 (b)* is excluded, as it is less realistic to have chronic pain (`ChP`) affect cannabidiol consumption (`CBD`). In our opinion, it is more sensible to have cannabidiol consumption (`CBD`) affect chronic pain (`ChP`).

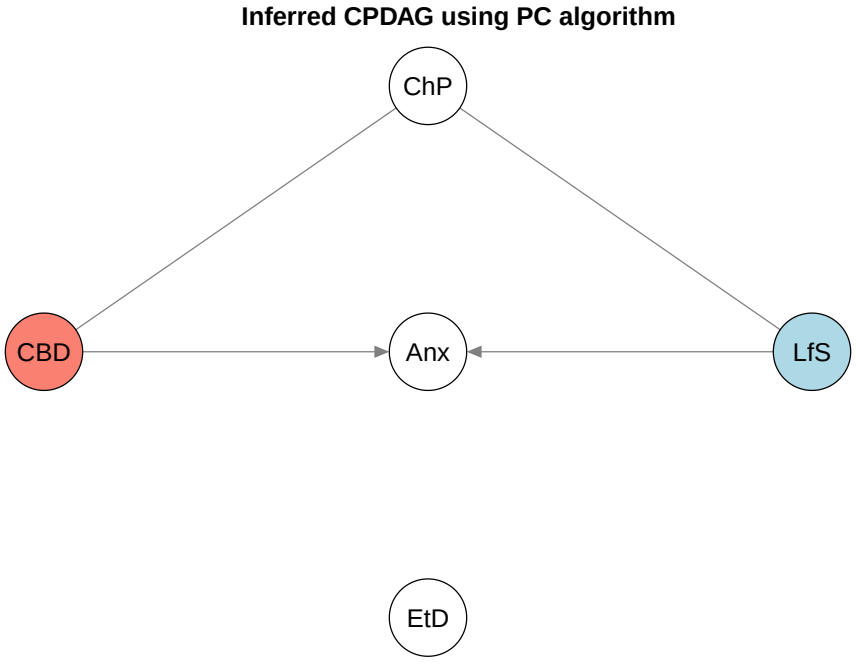

Figure 2: Completed Partially Directed Acyclic Graph (CPDAG)

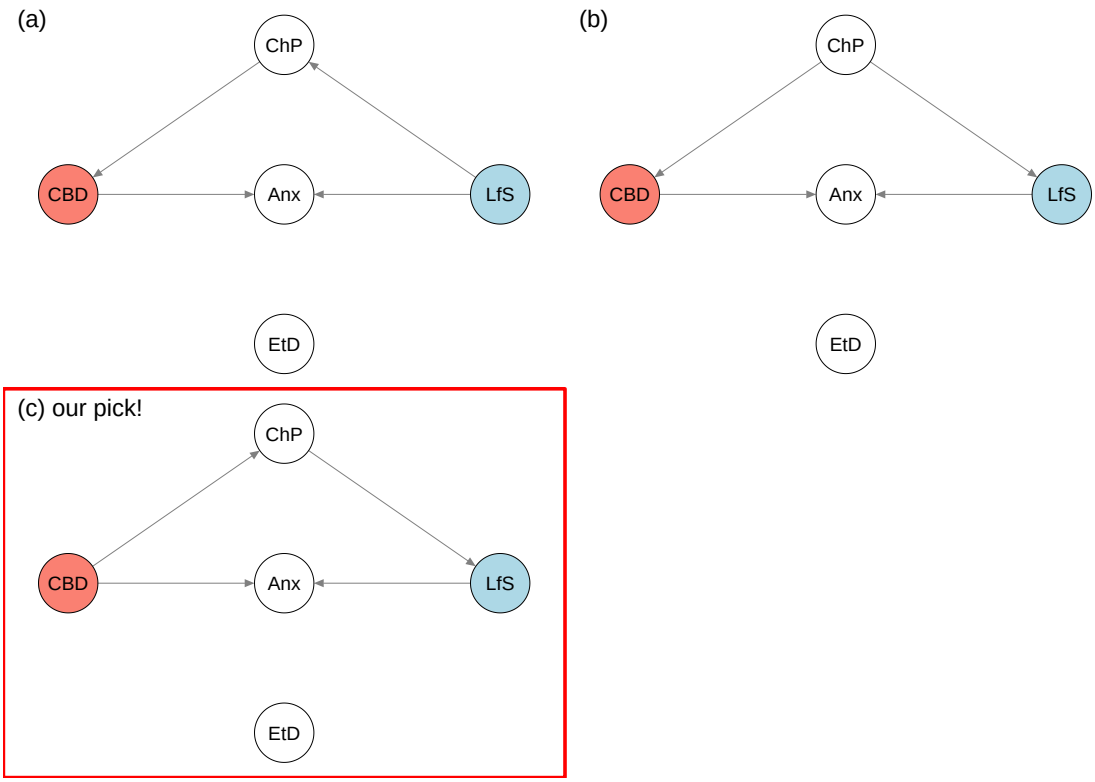

Figure 3: Estimated Markov-Equivalence Class

3. The other pair of students has specified for you which specific causal relationship you should estimate. What were the relevant cause and effect variables they specified for you?

a. Cause variable: `CBD`
b. Outcome variable: `LfS`

4. Estimate the prima facie effect for the causal effect of question 3. Discuss the results.

In order to estimate the prima facie effect, here we look at the continuous version: `CBD` and the dichotomized version: `CBDdichotomous`.

- The prima facie effect of `CBD` is estimated to be 0.11 when using the continuous version, and is statistically significant ($p < 0.01$). It indicates that the `LfS` (life satisfaction) is expected to increase by 0.11 with every milliliter increase in the cannabidiol consumption per day.
- The prima facie effect of `CBDdichotmous` is estimated to be 2.93 when using the dichotomous version, and is statistically significant ($p < 0.01$). It indicates that the `LfS` (life satisfaction) is expected to increase by 2.93 units when moving from the below average cannabidiol consumption to the above average cannabidiol consumption per day. We double-checked this by using the t-test and taking the estimated mean difference. This yields the same result, which

indicates that the average life satisfaction of people who consumed more than average of CBD (`CBDdichotmous` = 1) is 2.93 units higher than that of people who consumed less than average of CBD (`CBDdichotmous` = 0). See *Figure4* for the distribution of `LfS` (life satisfaction) for each group. (*Note*: we checked whether the dichotomization was done based on the mean value of CBD, and we were able to confirm that it was indeed mean-splitted. The code is available in the *Appendix*).

- The estimate of prima facie effect when using the continuous `CBD` differs from the estimate when using the `CBDdichotomous`. This is the case, because the cause variables are simply different. With the continuous variable, we are looking at the effect of every unit change on the original scale of `CBD`, whereas with the dichotmous variable, we are looking at the effect of going from below average to above average of CBD consumption.

```
## Using continuous CBD
mod.con <- lm(LfS ~ CBD , data = cannabis.dat)
round(summary(mod.con)$coefficients, 4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.938     1.8286 22.3877   0.0000
## CBD            0.106     0.0323  3.2810   0.0011
```

```
## Using dichotomized CBD
mod.dic <- lm(LfS ~ CBDdichotomous, data = cannabis.dat)
round(summary(mod.dic)$coefficients, 4)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     45.2622     0.6954 65.0861   0.0000
## CBDdichotomous   2.9306     0.9874  2.9679   0.0031
```

```
## Double-check with the t-test result
t.test1 <- t.test(LfS ~ CBDdichotomous, data=cannabis.dat)
# Difference in means between the groups:
m1 <- t.test1$estimate[[2]]
m0 <- t.test1$estimate[[1]]
m1 - m0
```

```
## [1] 2.93063
```

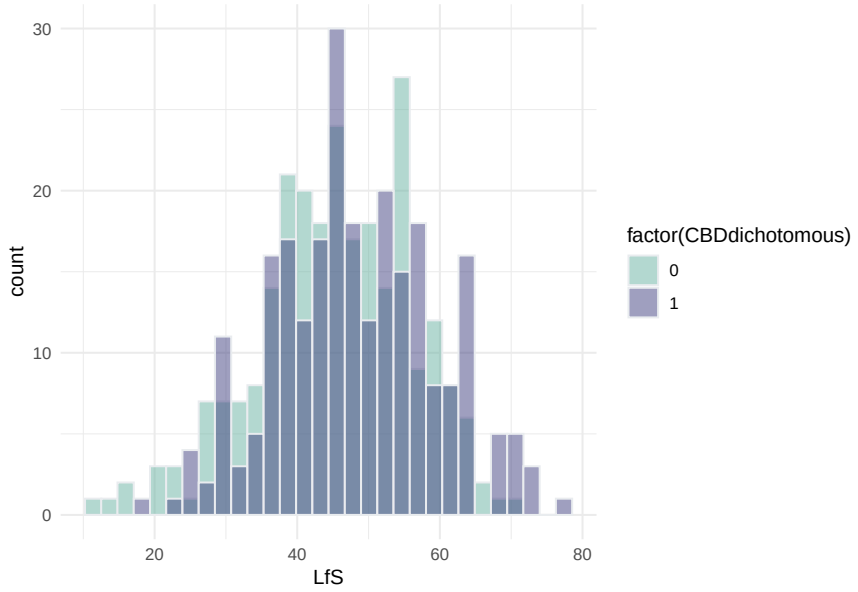

Figure 4: Histogram of Life Satisfaction per each CBD group

5. For the causal effect of question 3...

5a. Based on your selected DAG (2b), what linear regression model should be used to estimate the causal effect correctly and why?

According to the DAG chosen in 2b: *Figure3 (c)*, `ChP` is a mediator and should not be controlled for, since we are interested in the total effect of `CBD` on `LfS`. The path via `Anx` is blocked as `Anx` is a collider. Thus, we do not need to control for any variables.

The linear regression model to estimate the causal effect of `CBD` on `LfS` can, therefore, be written as:

$$LfS_i = \beta_0 + \beta_1 CBD_i + \epsilon_i, \text{ with } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \ \forall\, i = 1, \ldots, 500$$

5b. Estimate the causal effect with this regression model (with your continuous cause variable). Discuss the results.

Based on the regression model specified in 5a, the causal effect of `CBD` (continuous variable) on `LfS` is estimated to be 0.11 (SE= 0.03), and it is statistically significant ($p < 0.01$). This is the same as the estimate of prima facie effect in Q4, because we do not have to control for any variables. As there is no confounder, the prima facie effect in this case is essentially the same as the true causal effect. The interpretation is hence the same as described above: the `LfS` (life satisfaction) is expected to increase by 0.11 units with every milliliter increase in cannabidiol consumption per day.

```r
## Causal effect : Continuous CBD
mod.con <- lm(LfS ~ CBD, data = cannabis.dat)
round(summary(mod.con)$coefficients, 4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.938     1.8286 22.3877   0.0000
## CBD            0.106     0.0323  3.2810   0.0011
```

5c. Now estimate the causal effect with the dichotomous cause variable. Discuss the results.

The causal effect of `CBDdichotomous` (dichotomous variable) on `LfS` is estimated as 2.93 (SE= 0.99), and it is statistically significant ($p < 0.01$). Again, this is the same as the estimate of the prima facie effect in Q4, since there is no variable needed to be controlled for. Thus, the prima facie effect is actually a good estimator of the true causal effect in this particular case. Given that the dichotomization of `CBD` was based on mean splitting, we can interpret the result as `LfS` (life satisfaction) is expected to increase by 2.93 units when going from the below average cannabidiol consumption to above average cannabidiol consumption per day. As explained above, the estimate of causal effect with the dichotomous variable is different from the one estimated with the continuous variable, because we are now evaluating the effect of going from below average consumption to above average consumption of `CBD`, rather than looking at a unit change in `CBD` consumption as we did in 5b.

```r
## Causal effect : Dichotomous CBD
mod_dichotomized <- lm(LfS ~ CBDdichotomous  , data = cannabis.dat)
round(summary(mod_dichotomized)$coefficients, 4)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     45.2622     0.6954 65.0861   0.0000
## CBDdichotomous   2.9306     0.9874  2.9679   0.0031
```

6. For the causal effect of question 3, based on your selected DAG, use either matching, inverse propensity weighting, or subclassification to estimate the causal effect of your dichotomous cause variable on the outcome. Briefly discuss the steps you take and the results.

Disclaimer: Since we do not have any confounders based on our selected DAG, we could simply take that mean difference (e.g., prima facie effect) to estimate the average causal effect. Thus, we assume that this step of computing propensity score and trying to get both groups balanced across the covariates is not very relevant in our case. However, we will still present the results below based on the estimated propensity scores using all the available variables (`EtD`, `Anx`, `ChP`). We are interested in seeing if the chosen technique is successful in reducing the standardized mean difference for the other variables and to what extent this changes the ACE estimate.

Step1) Investigate the covariates

As shown below, the standardized mean differences (SMD) of `ChP` and `Anx` are larger than the rule of thumb: 0.1, implying that there is an imbalance between the two groups (higher CBD consumption / lower CBD consumption) with respect to `ChP` and `Anx`.

```
##                     Stratified by CBDdichotomous
##                      0              1              SMD
##   n                    252            248
##   EtD (mean (SD))     0.07 (7.99)    0.49 (7.66)    0.052
##   ChP (mean (SD))    47.94 (12.16)  39.34 (12.79)   0.688
##   Anx (mean (SD))    66.78 (13.03)  55.90 (13.50)   0.820
```

Step2) Estimating the propensity scores

We compute the propensity scores by running a logistic regression model in which the dichotomous version of CBD is the outcome variable and all covariates (`EtD`, `ChP` &`Anx`) are included as predictors.

Figure5 shows the distribution of propensity scores for the high and the low CBD consumption group. We can see that their distributions overlap quite well for the most part, except for both ends of the tails. For the propensity scores lower than around 0.15, there are only people from the low CBD consumption, whereas for the propensity scores higher than around 0.9, there are only people with high CBD consumption. Based on this, we conclude that the positivity assumption is violated to some extent. Nevertheless, we will go on with our analysis, but this should be kept in mind.

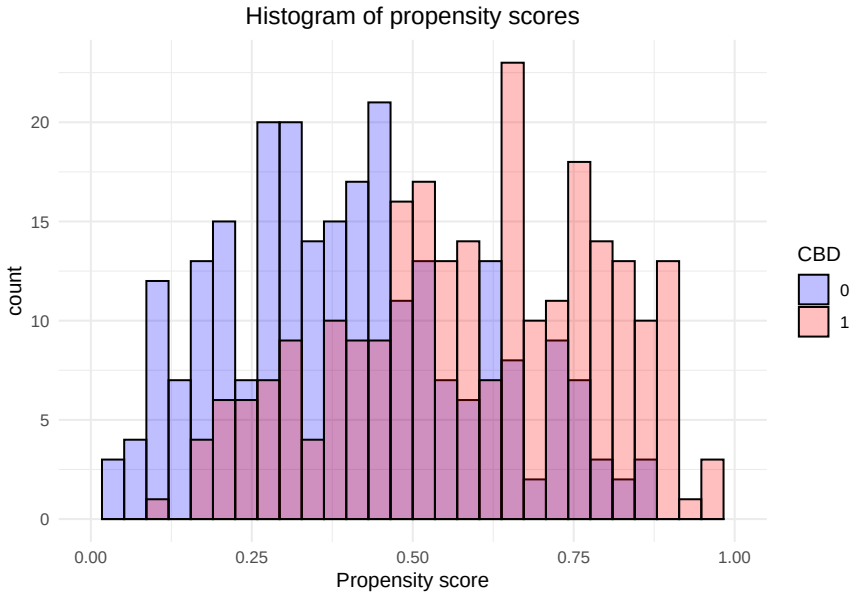

Figure 5: Distribtuion of propensity scores

Step3) Perform the analysis using the preferred method

We decide to use i) Inverse Propensity Weighting (IPW) and ii) Subclassication and compare the results from both methods.

Step 3-i) Inverse Propensity Weighting (IPW)

Here, we use the inverse of the estimated propensity scores as weights in order to correct for the over/under-representation in each group. The aim is to create a pseudo-population, where both groups are balanced across the covariates.

```
##                   Stratified by CBDdichotomous
##                    0               1               SMD
##   n                494.24          496.00
##   EtD (mean (SD))    0.58 (7.95)     0.15 (7.76)    0.054
##   ChP (mean (SD))   43.98 (12.63)   43.11 (12.42)   0.069
##   Anx (mean (SD))   61.98 (13.33)   61.10 (14.03)   0.064
```

```
msm <- svyglm(LfS ~ CBDdichotomous, design = weighteddata)
round(coef(summary(msm)), 4)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     47.4883     0.7392 64.2407   0.0000
## CBDdichotomous  -2.2157     1.0738 -2.0634   0.0396
```

As shown above, the standardized mean differences (SMD) are all smaller than 0.1 in the pseudo-population created by applying the inverse propensity weight. This indicates that the low- and the high-consumption groups do not differ much any more on the covariates and it is accordingly concluded that IPW is successful in mimicking an RCT.
Based on IPW procedure, the causal effect of CBD is estimated as -2.22 (SE=1.07), which is statistically significant ($p < 0.05$). It could be interpreted as the life satisfaction of people who consumed more than average of CBD is expected to be 2.22 units lower than that of people who consumed less than average of CBD. This is actually in contrast to the result we obtained previously in Q5 (estimate of causal effect of CBD = 2.93) and this point will be further discussed in detail in Q7.

Step 3-ii) Subclassification

We begin by creating five strata based on the propensity scores. We decide to create the strata in such a way that each stratum contains 20% of the data.
As shown in *Figure6*, the outer two strata are slightly wider. However, we do not consider them to be too wide and also considering the fact that the overlap is not so great in those ends, we decide to keep the strata as they are.

When looking at the SMD in the 2nd - 4th strata below, it can be seen that the SMDs are smaller than the cut-off, 0.1 (a couple of exceptions: `Anx` in 3rd stratum, EtD‘ in 4th stratum). However, in the outer strata (1st and 5th stratum), the SMD of all covariates are considerably higher than 0.1, which implies that there still exists a substantial imbalance between two groups within those strata. This can be related back to the violation of positivity shown in in *Figure6*.

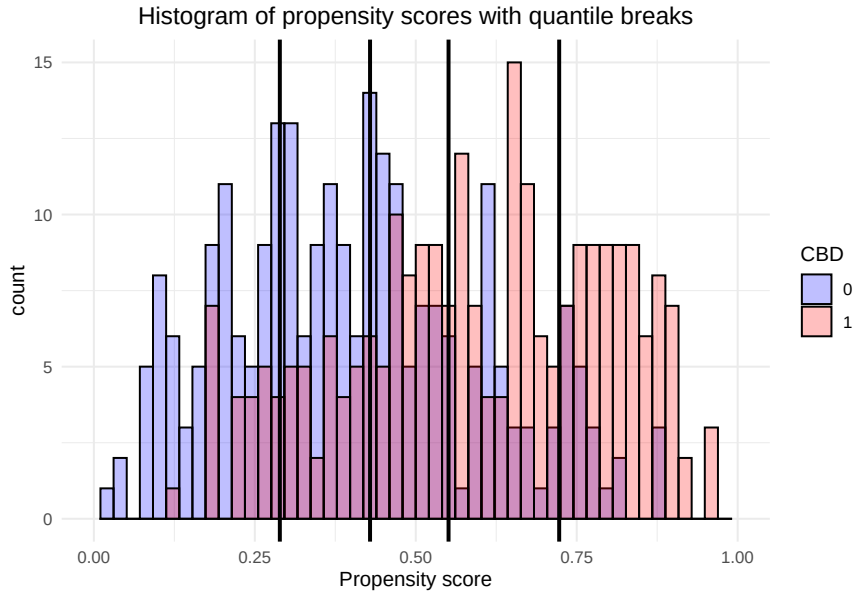

Figure 6: Distribution of propensity scores with strata

```
##                   Stratified by CBDdichotomous
##                    0             1             SMD
##   n                   79            21
##   EtD (mean (SD))   0.18 (8.39)   -0.14 (8.42)  0.038
##   ChP (mean (SD)) 58.11 (9.65)   54.05 (6.94)  0.483
##   Anx (mean (SD)) 79.24 (10.27) 76.92 (7.18)  0.262
##                   Stratified by CBDdichotomous
##                    0            1             SMD
##   n                   66           34
##   EtD (mean (SD)) -0.02 (8.68) -0.61 (8.34)  0.069
##   ChP (mean (SD)) 48.90 (8.37) 48.77 (8.33)  0.016
##   Anx (mean (SD)) 67.95 (7.38) 67.98 (6.15)  0.004
```

```
##                   Stratified by CBDdichotomous
##                    0             1             SMD
##   n                   51            49
##   EtD (mean (SD))   0.31 (7.84)   1.06 (7.58)  0.098
##   ChP (mean (SD)) 44.12 (8.62) 44.42 (8.45)  0.035
##   Anx (mean (SD)) 62.20 (6.48) 61.13 (7.97)  0.148
##                   Stratified by CBDdichotomous
##                    0             1             SMD
##   n                   34            66
##   EtD (mean (SD)) -2.14 (5.68)   0.07 (7.92)   0.320
##   ChP (mean (SD)) 38.38 (8.33) 38.32 (10.12)  0.007
##   Anx (mean (SD)) 54.99 (6.33) 54.52 (8.61)   0.063
##                   Stratified by CBDdichotomous
##                    0             1             SMD
##   n                   22            78
##   EtD (mean (SD))   2.85 (7.37)   1.12 (7.07)   0.239
##   ChP (mean (SD)) 32.12 (8.99) 28.96 (11.28)  0.310
##   Anx (mean (SD)) 47.38 (6.92) 42.86 (8.77)   0.572
```

We proceed by performing a t-test in each of the strata separately and take the difference in means between two groups (high and low CBD consumption). As seen below, the mean differences vary quit a bit across the strata, and only in the stratum 2, 3, and 4, we find a significant difference. In stratum 1 and 5, the difference is turned out to be not significant, but we can again presume that this is probably due to the lack of overlapping in both tails as we have seen above in *Figure6*.
The average causal effect is estimated as -2.01. We can interpret it in the same ways as we described above: the life satisfaction is expected to decrease by 2.01 units when going from the below average cannabidiol consumption to the above average cannabidiol consumption per day. This result is comparable to the one we obtained by using the IPW method. Therefore, we conclude that our results are quite robust as two different methods lead to the more or less same conclusion: the estimated causal effect of CBD on `LfS` (life satisfaction) is negative (appx. -2).

```
##          mean difference p-value
## stratum1          -1.649   0.465
## stratum2          -3.896   0.047
## stratum3          -3.600   0.024
## stratum4          -3.992   0.027
## stratum5           3.078   0.174
```

```
## average mean difference = -2.011821
```

7. What is your overall conclusion about the causal effect based on your results/reflections from steps 2-6?

The PC algorithm provided three possible DAGs for the given data set. The DAG we chose has a rather simple structure, where only four variables are related and the cause variable of interest (`CBD`) has only one valid causal path to the outcome variable of interest (`LfS`) via a mediator `ChP`. We have some doubts about the structure of this chosen DAG, especially regarding the direction between `Anx` and `LfS` (`Anx` ← `LfS`), but among the possible choices, this DAG matches best with our expected DAG structure (*Figure1*).

Based on the selected DAG, we do not have to control for any variables, and correspondingly it enables us to take the estimate of the prima-facie effect as our estimate of the true causal effect. The estimated causal effects are 0.11 and 2.93 for the continuous version of `CBD` and the dichotomized version of `CBD`, respectively.

Practically, if we were to believe in this DAG structure, we would draw the conclusion at this point, saying that our estimated true causal effects are the aforementioned values, as there is no confounder according to our DAG. Yet, we continue to investigate further, whether the two groups (high- and low CBD consumption) are indeed balanced across the other variables It is observed that the groups actually differ quite a bit on the variables `ChP` and `Anx`. Hence, we proceed with performing the analysis using IPW and Subclassification method to see if the estimate of causal effect would differ from what we saw initially without controlling for any variables.

The results show that when controlling for the other variables, the causal effect of `CBD` is actually quite opposite from the estimates of prima-facie effect, as one would expect. The estimated causal effects from IPW and Subclassification are -2.22 and -2.01, respectively, which indicates that the life satisfaction of people with above average CBD assumption is expected to be lower (by around 2 units) than that of the people with below average CBD assumption. This is indeed in contrast to the estimate of prima-facie effect, where we say life satisfaction is expected to increase by 2.93, when going from the below average CBD consumption to the above average CBD consumption. We find it very interesting that the sign flips (positive → negative effect) when we perform the analyses using propensity scores. We assume that this is actually an example of Berkson's bias, where you control for a collider and get a biased estimate of the causal effect. `Anx` is a collider in our DAG, but we ignored that and used all variables including `Anx` to compute the propensity scores. Hence, we got a different result, in fact, a totally opposite result from what we initially estimated in Q5.

We did not include the code here, but we tried modelling the propensity scores with different predictors: only `ChP`, only `EtD`, only `Anx`, `ChP` + `EtD`, and `Anx` + `ChP`. When we included `ChP` as a predictor, the result showed that the effect of `CBD` was estimated to be smaller than the original estimate from Q5 and non-significant. This makes sense, since `ChP` is a mediator according to our DAG, hence controlling for it would make the causal effect of CBD smaller (if not 0, as theoretically this is the only valid path as per our DAG). Additionally, when we had `EtD` as the predictor, the result was almost the same as the original estimate. This again makes sense, since `EtD` is not related to any variables in our DAG, so theoretically controlling for it should not make any difference. And again, controlling for only `Anx` drastically changed the causal effect from positive to negative, and as we explained above, this is likely due to the fact that `Anx` is a collider according to our DAG. All in all,

the results turn out to be in accordance with the chosen DAG structure. The most intriguing thing is seeing the sign flips when `Anx` is controlled for, which is the variable that we initially found quite odd to be a collider from a substantive perspective.

Our overall conclusion is that we are not completely sure what would be a correct conclusion in this case. According to everything we did above, we would say that the causal effect of `CBD` on `LfS` (life satisfaction) is 0.11 for the continuous case, and 2.93 for the dichotomized case. However, due to the uncertainty in our DAG structure, we are not entirely convinced that this is the estimate of the true causal effect. Not only that we have found literature that contradicts the DAG structure (e.g., `Anx` $\rightarrow$ `LfS`; see Q1), but also we think that there is not enough power to detect the correct DAG in this case, considering that the sample size is only 500. We believe that this is a much-simplified version of DAG, yet even then we think that there needs to be more data/information in order to get a better (i.e., closer to truth) estimate of the causal effect of `CBD` on `LfS`.

8. What do you take away about causal inference from this and the previous assignment (if anything)?

Our main take away is: Causal inference is difficult! We realized the first assignment was rather simple, now that we got into the situation of trying to estimate the causal effect where we had no idea what was going on (2nd assignment). But we understand now that this is closer to the real case of analysis, which we would be likely to face in the future.

Another thing we realized is that the choice of what is assumed to be the correct DAG is very subjective. Although literature can be found to reason for or against some directed relationships, it is still up to the analyst to choose the DAG in the end. It showed us that for causal inference it is of greater importance to have some background knowledge in the field of research and to exchange ideas with the other experts.

References

Aggarwal, S. K., Carter, G. T., Sullivan, M. D., Zumbrunnen, C., Morrill, R., & Mayer, J. D. (2013). Prospectively Surveying Health-Related Quality of Life and Symptom Relief in a Lot-Based Sample of Medical Cannabis-Using Patients in Urban Washington State Reveals Managed Chronic Illness and Debility. *American Journal of Hospice and Palliative Medicine*, *30*(6), 523–531. https://doi.org/10.1177/1049909112454215

Alzaher, W. (2022). How Can CBD Help With Anorexia?. *EATING DISORDER, MEDICAL CONDITIONS*. https://cannabisclinic.co.nz/medical-conditions/.

Blessing, E.M., Steenkamp, M.M., Manzanares, J., & Marmar, C.R. (2015). Cannabidiol as a Potential Treatment for Anxiety Disorders. *Neurotherapeutics*, *12*, 825–836. https://doi.org/10.1007/s13311-015-0387-1

Claydon, E.A., DeFazio, C., Lilly, C.L., & Zullig, K.J. (2020). Life satisfaction among a clinical eating disorder population. *Journal of Eat Disorder*, 8(53). https://doi.org/10.1186/s40337-020-00326-z

Hill, K.P. (2015). Medical Marijuana for Treatment of Chronic Pain and Other Medical and Psychiatric Problems: A Clinical Review. *JAMA*. *313*(24):2474–2483. https://doi.org/10.1001/jama.2015.6199.

Serin, N.B., Serin, O., & Özbaş, L.F. (2010). Predicting university students' life satisfaction by their anxiety and depression level, *Procedia - Social and Behavioral Sciences*, *9*, 579-582. https://doi.org/10.1016/j.sbspro.2010.12.200.

Appendix

Code for Figure1

```
## plot the DAG we expect using 'dagify'
cannabis <- dagify(
  ChP ~ CBD,
  Anx ~ CBD,
  EtD ~ Anx,
  LfS ~ ChP + Anx + EtD,
  exposure = "CBD", # cause variable we are interested in
  outcome = "LfS", # effect variable we are interested in
  #set the coordinates
  coords = list(x = c(CBD = -1, ChP= 0, Anx= 0, EtD = 0, LfS = 1),
                y = c(CBD = 0, ChP=1, Anx=0, EtD = -1, LfS = 0))
)

ggdag_status(cannabis) + theme_dag()
```

Code for PC-algorithm (Figure2)

```
## plot the markov equivalence class
# extract the adjacency matrix of the cpdag
cpdag_mat <- as(pc_fit, "matrix")

# extract the DAG adjacency matrix in a vector form (by rows)
res <- pdag2allDags(cpdag_mat)

# get the adjacency matrix of an individual DAG
res_dags <- list()
for(i in 1 :nrow(res$dags)){
  res_dags[[i]] <- t(matrix(res$dags[i,], 5, 5, byrow = T))
}

# specify a layout matrix
laymat <- rbind(c(-1, 0),
                c(0, 0),
                c(0, -1),
                c(0, 1),
                c(1, 0))
```

```r
## plot the CP-DAG
# adjust the plotting margin
par(mar=c(5, 4, 3.5, 2) + 0.1)

# plot the CP-DAG using qgraph
qgraph(pc_fit, color=c("salmon", "white", "white", "white", "lightblue"),
       layout = laymat, labels = varnames)
title("Inferred CPDAG using PC algorithm", line=3, cex.main = 1)
```

Code for Figure3

```r
# plot the DAG for each adj.matrix
par(mfrow=c(2,2))
  qgraph(res_dags[[1]], bidirectional=TRUE, color=c("salmon", "white",
                                                    "white", "white", "lightblue"),
       layout = laymat, labels = varnames, asize = 5, vsize =14,
       title = "(a)")

    qgraph(res_dags[[2]], bidirectional=TRUE, color=c("salmon", "white",
                                                    "white", "white", "lightblue"),
       layout = laymat, labels = varnames, asize = 5, vsize =14,
       title = "(b)")

    qgraph(res_dags[[3]], bidirectional=TRUE, color=c("salmon", "white",
                                                    "white", "white", "lightblue"),
       layout = laymat, labels = varnames, asize = 5, vsize =14,
       title = paste0("(",letters[i],")"," ", "our pick!"))
  box("figure", col="red", lwd=2)
```

Code for Figure4

```r
cannabis.dat %>%
ggplot(aes(x = LfS, fill = factor(CBDdichotomous))) +
  geom_histogram( color="#e9ecef", alpha = 0.5, position = 'identity') +
  scale_fill_manual(values=c("#69b3a2", "#404080")) +
  theme_minimal()
```

Code for checking if the dichomoization is based on the mean-split

```
## Checking whether it is mean-splited
table(cannabis.dat$CBDdichotomous)
high <- which(cannabis.dat$CBD > mean(cannabis.dat$CBD))
low <- which(cannabis.dat$CBD < mean(cannabis.dat$CBD))

## they are the same --> it is indeed mean-splitted
cannabis.dat$CBDdichotomous[high] == cannabis.dat$
  CBDdichotomous[which(cannabis.dat$CBDdichotomous==1)]
cannabis.dat$CBDdichotomous[low] == cannabis.dat$
  CBDdichotomous[which(cannabis.dat$CBDdichotomous==0)]
```

Code for computing the propensity scores

```
## Computing the propensity scores
# Run the logistic regression
logreg <- glm(CBDdichotomous ~ EtD+ChP+Anx, family="binomial", data=cannabis.dat)
# Obtain the prediction of the probability of CBDdichotomous = 1
ps <- predict(logreg, type = "response")
# Add this predicted probability to the data
cannabis.dat$ps <- ps
```

Code for IPW analysis

```
## IPW procedure
# assign the weight (inverse of propensity scores) for each group
weight <- ifelse(cannabis.dat$CBDdichotomous==1, 1/(cannabis.dat$ps), 1/(1-cannabis.dat$ps))
# create a pseudo population
weighteddata <- svydesign(ids = ~ 1, data =cannabis.dat, weights = ~ weight)
# check the SMD
weightedtable <- svyCreateTableOne(vars = c("EtD", "ChP", "Anx"),
                                   strata = "CBDdichotomous",
                                   data = weighteddata, test = FALSE)
print(weightedtable, smd = TRUE)
```

Code for Figure5

```
# distributon of propensity scores
cannabis.dat %>%
  ggplot(aes(x=ps, fill=factor(CBDdichotomous))) +
  geom_histogram(color="black", position="identity", bins=30) +
  scale_fill_manual(values=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)), name="CBD") +
  labs(title="Histogram of propensity scores", x = "Propensity score") + xlim(0,1) +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```

Code for Figure6

```
# dividing it to 5 strata
cannabis.dat$stratum <- cut(cannabis.dat$ps,
                    breaks=c(quantile(cannabis.dat$ps, probs=seq(0,1,0.2))),
                    labels=seq(1:5),
                    include.lowest=TRUE)

br <- c(quantile(cannabis.dat$ps, probs=seq(0,1,0.2)))

# histogram of subclassification
cannabis.dat %>%
  ggplot(aes(x=ps, fill=factor(CBDdichotomous))) +
  geom_histogram(color="black", position="identity", bins=50) +
  scale_fill_manual(values=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)), name="CBD") +
  labs(title="Histogram of propensity scores with quantile breaks",
       x = "Propensity score") + xlim(0,1) + theme(plot.title = element_text(hjust = 0.5))+
  geom_vline(aes(xintercept = br[2]), size=1)+
  geom_vline(aes(xintercept = br[3]), size=1)+
  geom_vline(aes(xintercept = br[4]), size=1)+
  geom_vline(aes(xintercept = br[5]), size=1)+
  theme_minimal() + theme(plot.title = element_text(hjust = 0.5))
```

Code for t-tests in Stratification

```
# Perform a t-test in each stratum
results <- matrix(NA, 5,2)
rownames(results) <- c("stratum1", "stratum2", "stratum3", "stratum4", "stratum5")
colnames(results) <- c("mean difference", "p-value")

for (quantiles in 1:5) {
  t.test3 <- t.test(LfS ~ CBDdichotomous, data = cannabis.dat[which(
    cannabis.dat$stratum==quantiles),])
  print(t.test3)
  # Difference in means:
  results[quantiles,1] <- t.test3$estimate[[2]] - t.test3$estimate[[1]]
  # p - value:
  results[quantiles,2] <- t.test3$p.value
}
```