MY457: Problem Set 1

L. Cäcilia Präckel

Mi/31/Jan

1 Concepts

1.1 The notation Y_i1 is used for the potential outcome of the response variable Y_i when the explaining variable D is set to 1. This means, we look at an unobserved, hypothetical outcome that is counterfactual to Y_i0. For the same unit i, the quantity of Y_i1 would be equal to the observed value Y_i if Y_i1 has been realised due to the treatment effect (D set to 1).

1.2 $E[Y_i0|D_i = 1]$ is the sum of all potential outcomes of Y_i0 under the condition of D=1, meaning the treatment effect is switched on. $E[Y_i0|D_i = 0]$ is the sum of all possible outcomes of Y_i0 under the condition of D=0, meaning the treatment effect is switched off. These quantities should be equal when the treatment effect is 0 because apparently, the condition of D_i is non-informative, D and Y are not associated. The quantities should be unequal any time the treatment effect is bigger than 0, because then it has an affect on the two variables.

1.3 The average treatment effect ATE is defined as the mean of the differences of individual potential outcomes (in a binary case, the difference between Y_i0 and Y_i1). If we randomly assign individuals to the treatment and control group, we can identify the ATE because the assignment method is not dependent on any confounding variable X. We could also calculate the ATE when having non-randomly assigned treatment and control groups but then the calculated ATE might differ from the identified ATE.

2 Simulations

2.1 We first set a seed to enable reproducibility of the random numbers. We then define the number of observation to be 500 and the ATE to be 5000. A dataframe is created with three columns, age, being normally distributed, education, which is a categorical variable of 1,2,3 or 4 for each observation, and Y0, which is the potential outcome without treatment, also normally distributed.

With mutate, we then add three new columns, one being the potential outcome Y1, which is calculated by adding the treatment effect to the potential outcome of Y0, one being D the explaining variable, which takes the randomly assigned values 1 or 0 for each observation, one being Y, the outcome variable which takes the value Y1 if D equals 1, otherwise Y0.

2.2

```
##
##  Welch Two Sample t-test
##
## data:  data$Age by D
## t = -1.3778, df = 485.8, p-value = 0.1689
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.9178207  0.5124662
## sample estimates:
## mean in group 0 mean in group 1
##        41.77824        42.98092
```

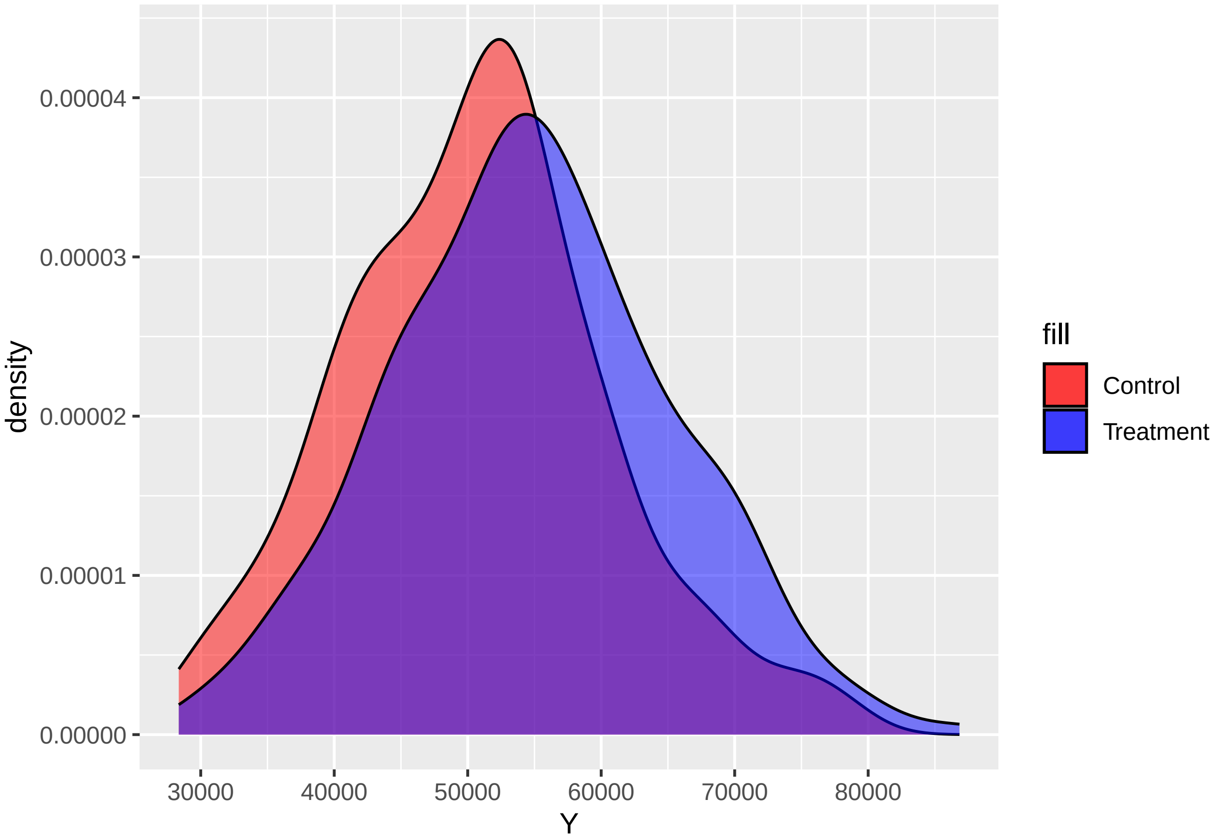
```
##
##  Welch Two Sample t-test
##
## data:  data$Education by D
## t = -0.53554, df = 492.69, p-value = 0.5925
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.2466875  0.1410121
## sample estimates:
## mean in group 0 mean in group 1
##        2.481061        2.533898
```

The mean of the potential outcomes for both age and education are close to each other, with 41.77 and 42.98, and 2.48 and 2.53, respectively. The two characteristics of the people in the experiment seem reasonably balanced.

2.3

```
## [1] "Difference-in-means of the potential outcomes: 5000"
```

The difference-in-means of the potential outcomes is equal to the average treatment effect. This is not surprising because we do not have any confounding variables, the data set is rather balanced, and treatment was assigned randomly. This is of course barely reproducible in a real-world environment.



2.4

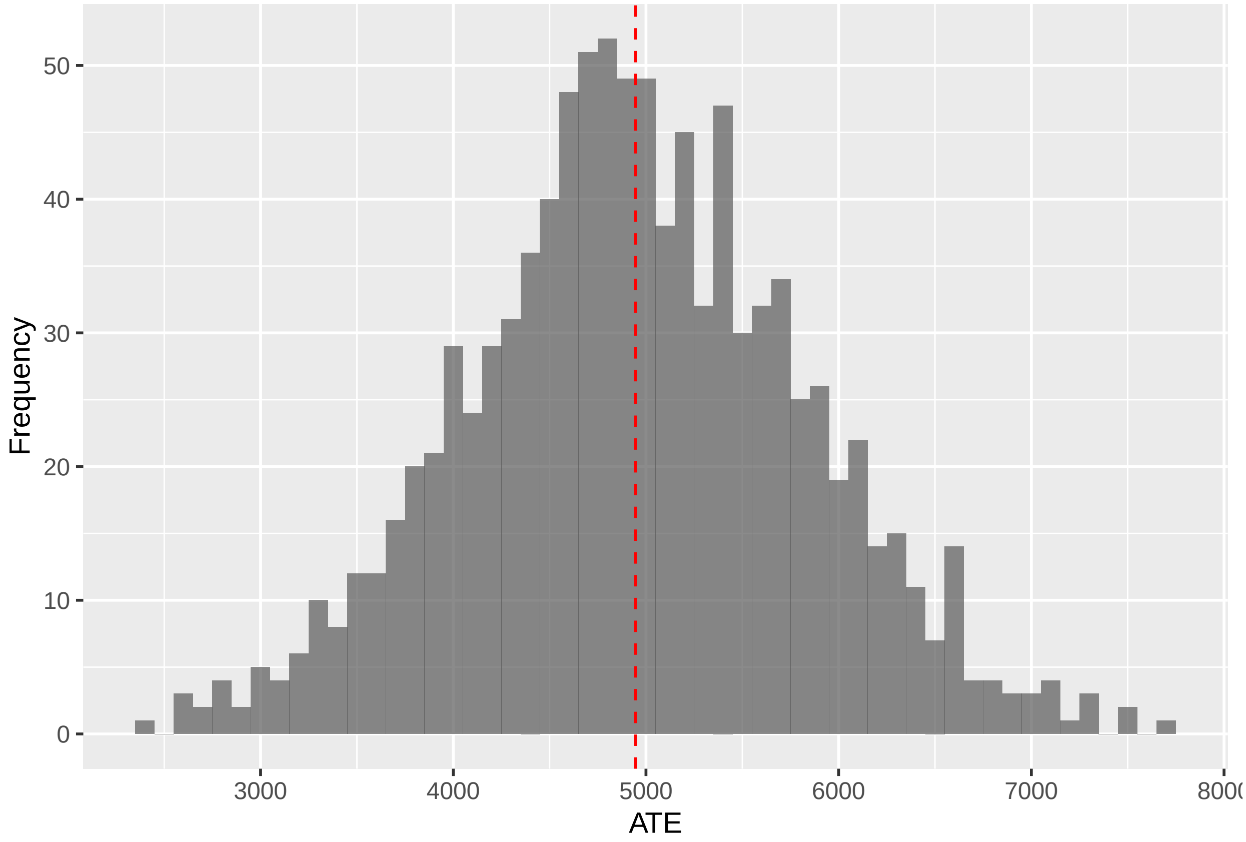
We can see from this plot that the treatment group has a lower maximum and the curve is less skewed.

```
## [1] "Difference-in-means between realised Y (treatment and control): 4181.02"
```

```
##
## Call:
## lm(formula = Y ~ D, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25286  -7123    -52   6148  32047
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  50612.0      625.2  80.947 < 0.0000000000000002 ***
## D             4181.0      910.1   4.594           0.00000551 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10160 on 498 degrees of freedom
## Multiple R-squared:  0.04066,    Adjusted R-squared:  0.03873
## F-statistic: 21.11 on 1 and 498 DF,  p-value: 0.000005514
```

Both possibilities, calculating the ATE with difference-in-mean and with a linear regression come to the same value of 4181.0. This is not very close to the ground truth of 5000, which we calculated earlier. We can see that the realised values of Y show a similar but not the same average treatment effect. This might be due to the slight imbalance in the age and education characteristic or is an coincidence of the distribution which resulted from this specific randomization. We could test this by running the experiment with increasing n.

2.5



The histogram shows that the ATEs are normally distributed around the true ATE of 5000. The mean of the ATEs is 4999.9, very close to the true ATE. This is not surprising because we now have a large sample size and the data is balanced. Now the ATE of 4181 from the linear regression and the t-test seems like a bad coincidence or a coding error. The ATEs are not exactly normally distributed, but the distribution is very close to normal.

3 Replication on the how_to_elect_more_women dataset

3.1 Creating a dummy variable which indictes if at least one women has been elected

3.2 Show the proportion of observations in each treatment/control group

```
## # A tibble: 5 x 2
##   condition              n
##   <dbl+lbl>          <int>
## 1  1 [Control]         541
## 2  2 [Supply]          539
## 3  3 [Demand]          538
## 4  4 [Supply+Demand]   538
## 5 NA                    11
```

The proportions in the treatment/control groups are quite balanced, ranging from 538 to 541 each.

3.3 Check if the two pre-trained characteristics age and distance from convention site are balanced

```
## [1] "The p-value of control vs other groups is  0.457497800595462"

## [1] "The p-value of treatment supply vs other groups is  0.62038205453933"

## [1] "The p-value of treatment demand vs other groups is  0.94905750372614"

## [1] "The p-value of treatment supply+demand vs other groups is  0.854952873013779"

## [1] "The p-value of control vs other groups is  0.352555206395007"

## [1] "The p-value of treatment supply vs other groups is  0.429815265752043"

## [1] "The p-value of treatment demand vs other groups is  0.622030110474536"

## [1] "The p-value of treatment supply+demand vs other groups is  0.645120978968075"
```

We can see from the comparison that for each treatment/control group the pre-treatment characteristics race and age are very balanced, with small p-values for each group.

3.4 Calculating the ATEs of different treatments with linear regression

```
##
## Call:
## lm(formula = sd_onefem2014 ~ condition, data = vote_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4592 -0.4359 -0.3891  0.5641  0.6109
```

```
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.36578    0.02838  12.890 <0.0000000000000002 ***
## condition    0.02336    0.01042   2.242              0.0251 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4938 on 1810 degrees of freedom
##   (355 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.002769,   Adjusted R-squared:  0.002218
## F-statistic: 5.026 on 1 and 1810 DF,  p-value: 0.02509
```