Computer Assignment 2a Costello

Rebecca Costello

9/3/2019

Rebecca Costello - Ties van der Veen - Niels van Opstal

```
#To improve the layout of the command summarise
knit_print.data.frame <- lemon_print
```

```
#to tell summarytools that we are working in R Markdown
st_options(plain.ascii = FALSE, style = "rmarkdown")
st_css()
```

```
## <style type="text/css">
##  img {   background-color: transparent;   border: 0; }  .st-table td, .st-table th {   padding: 8px;
```

```
theUrl_ca2a <- "https://surfdrive.surf.nl/files/index.php/s/ULZJ0bBbphCttpG/download"
library(haven)
students <- read_dta ("ca2a_2019.dta")
```

II. Potential outcomes

(a) Define the two potential outcomes for a particular student, Y(0,i) and Y(1,i). Y(0,i) = Risk perceptions remain the same T(1,i) = Risk perceptions changes.

(b) Thinking of a particular individual, why exactly would someone's perception of bicycle theft be affected by having been reminded of the last instance of bicycle theft? Their perception of bicycle theft could be affected because they were reminded of the last time this happened. This makes the memory more salient, which makes them consider the risk more readily. When someone first recalls the last instance of bicycle theft, they remenber the feelings/emotions attached to that experience. It is brought to the forfront of their mind and they use this information in answering the next question. This can also be called the availability heuristic.

III. Descriptive statistics

```
#overview of variables in the data set
summary(students)
```

```
##   frequentuser    bicyclestolen_ever     female       international
##  Min.   :0.0000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000     Median :0.0000   Median :0.0000
##  Mean   :0.8913   Mean   :0.3315     Mean   :0.3913   Mean   :0.4891
##  3rd Qu.:1.0000   3rd Qu.:1.0000     3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000
##  moved_notrecent    treatment     perception_person_low     age20
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000        Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000        1st Qu.:0.00000
##  Median :0.0000   Median :1.000   Median :0.0000        Median :0.00000
##  Mean   :0.2663   Mean   :0.538   Mean   :0.3207        Mean   :0.03261
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000        3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000        Max.   :1.00000
##    cohort2019
##  Min.   :0.0000
##  1st Qu.:0.0000
```

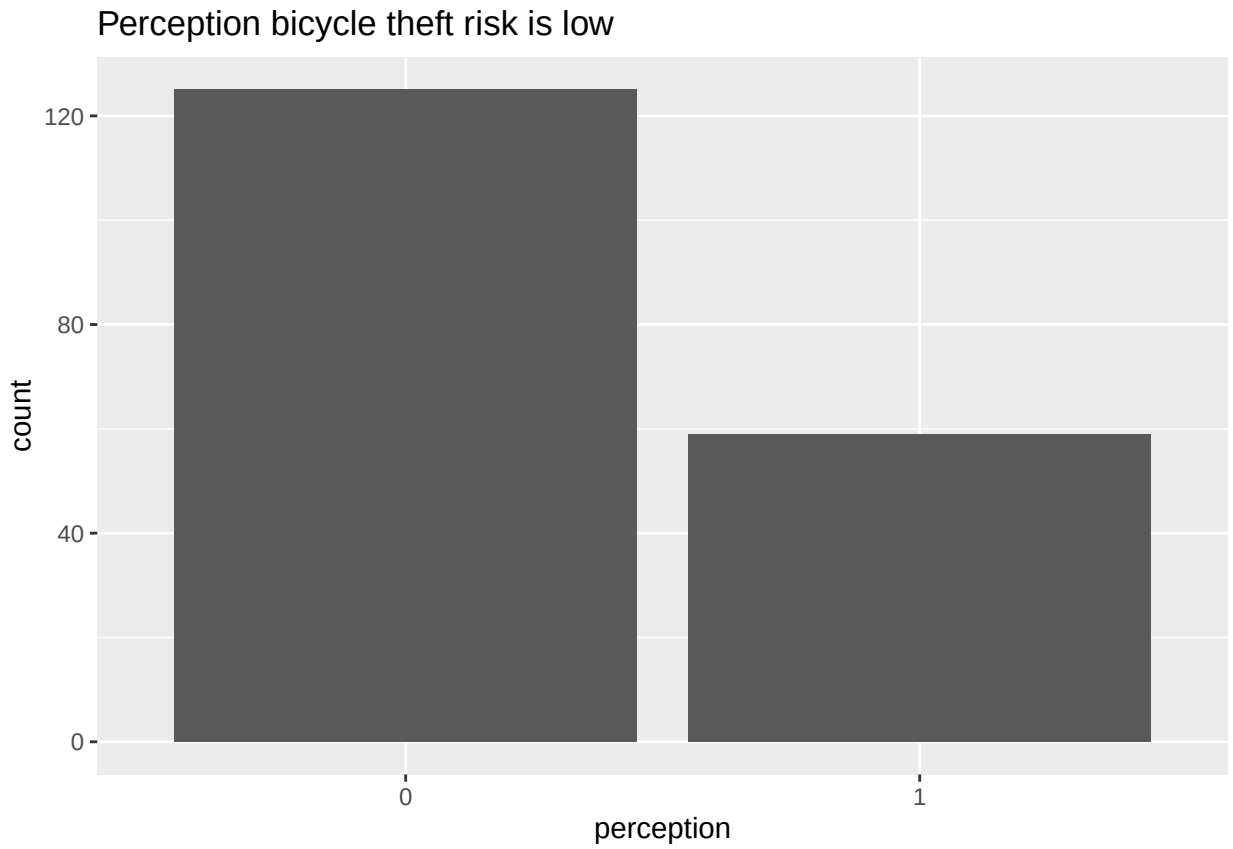
```
##  Median :0.0000
##  Mean   :0.3533
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
#summary of one variable in the data set
summary(students$perception_person_low)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3207  1.0000  1.0000
```

```
ggplot(students, aes(x=as.factor(perception_person_low)))+
geom_histogram(stat="count")+
labs(x='perception', y='count', title='Perception bicycle theft risk is low')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
#to know if there are any missings NA in the outcome variable
summary(is.na(students$perception_person_low))
```

```
##    Mode   FALSE
## logical     184
```

IV. Balance check

```
library(Matrix)
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
```

```
##
##     expand
```

```
#create a cross tab for females by treatment status
ctable(students$treatment, students$female)
```

```
## ### Cross-Tabulation, Row Proportions
## #### treatment * female
## **Data Frame:** students
##
## |           |        |             |            |              |
## |----------:|-------:|------------:|-----------:|-------------:|
## |           | female |           0 |          1 |        Total |
## | treatment |        |             |            |              |
## |         0 |        |  49 (57.6%) | 36 (42.4%) |  85 (100.0%) |
## |         1 |        |  63 (63.6%) | 36 (36.4%) |  99 (100.0%) |
## |     Total |        | 112 (60.9%) | 72 (39.1%) | 184 (100.0%) |
```

```
#cross tab for international
ctable(students$treatment, students$international)
```

```
## ### Cross-Tabulation, Row Proportions
## #### treatment * international
## **Data Frame:** students
##
## |           |               |             |             |              |
## |----------:|--------------:|------------:|------------:|-------------:|
## |           | international |           0 |           1 |        Total |
## | treatment |               |             |             |              |
## |         0 |               | 42 (49.4%) | 43 (50.6%) |  85 (100.0%) |
## |         1 |               | 52 (52.5%) | 47 (47.5%) |  99 (100.0%) |
## |     Total |               | 94 (51.1%) | 90 (48.9%) | 184 (100.0%) |
```

```
#cross tab for moved notrecent
ctable(students$treatment, students$moved_notrecent)
```

```
## ### Cross-Tabulation, Row Proportions
## #### treatment * moved_notrecent
## **Data Frame:** students
##
## |           |                 |             |             |              |
## |----------:|----------------:|------------:|------------:|-------------:|
## |           | moved_notrecent |           0 |           1 |        Total |
## | treatment |                 |             |             |              |
## |         0 |                 |  62 (72.9%) | 23 (27.1%) |  85 (100.0%) |
## |         1 |                 |  73 (73.7%) | 26 (26.3%) |  99 (100.0%) |
## |     Total |                 | 135 (73.4%) | 49 (26.6%) | 184 (100.0%) |
```

```
#cross tab for age20
ctable(students$treatment, students$age20)
```

```
## ### Cross-Tabulation, Row Proportions
## #### treatment * age20
## **Data Frame:** students
##
## |           |       |             |          |              |
## |----------:|------:|------------:|---------:|-------------:|
```

```
## |           | age20 |           0 |        1 |        Total |
## | treatment |       |             |          |              |
## |         0 |       |  82 (96.5%) | 3 (3.5%) |  85 (100.0%) |
## |         1 |       |  96 (97.0%) | 3 (3.0%) |  99 (100.0%) |
## |     Total |       | 178 (96.7%) | 6 (3.3%) | 184 (100.0%) |
```

```
#t test for females with and without treatment
t.test(students$female~students$treatment)
```

```
##
##  Welch Two Sample t-test
##
## data:  students$female by students$treatment
## t = 0.8252, df = 176.23, p-value = 0.4104
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0833447  0.2031308
## sample estimates:
## mean in group 0 mean in group 1
##       0.4235294       0.3636364
```

Difference between standard deviation of a variable and the standard deviation of the mean of that variable (standard error): Standard deviation (SD): measures the amount of variablity, or dispersion, for a subject set of data from the mean. Standard error (SEM): measures how far the sample mean of the data is likely to be from the population mean. SEM<SD

V. Statistical power effect size (d) - treatment effect (difference between means) divided by the standard deviation significance level (sig. level) level of power (power)

```
#mean and standard deviation of the outcome variable in control
students %>%
  filter(treatment==0) %>%
  summarise(mean=mean(perception_person_low), sd=sd(perception_person_low))
```

mean	sd
0.4	0.4928054

A 10 percentage-points change in the share of students with low perception of the bicycle theft risk is the minimum effect of this treatment. Calculate the treatment effect: $0.1/0.5 = 0.2$

```
#report the required sample size
pwr.t.test(n = NULL, d = 0.2, sig.level = 0.05, power = 0.8,
type = c("two.sample"), alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 393.4057
##               d = 0.2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr.t.test(n = 92, d = 0.2, sig.level = 0.05, power = NULL,
  type = c("two.sample"), alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 92
##               d = 0.2
##       sig.level = 0.05
##           power = 0.2711829
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

VI. Estimating treatment effect in a randomized trial

(a) Take a first peek at the treatment effect with a bar graph. First assign the means for control and treatment to a variable:

```
#assign the means for control and treatment to a variable
students_peek <- students %>% group_by(treatment) %>%
summarise(perception_person_low_mean=mean(perception_person_low))
```

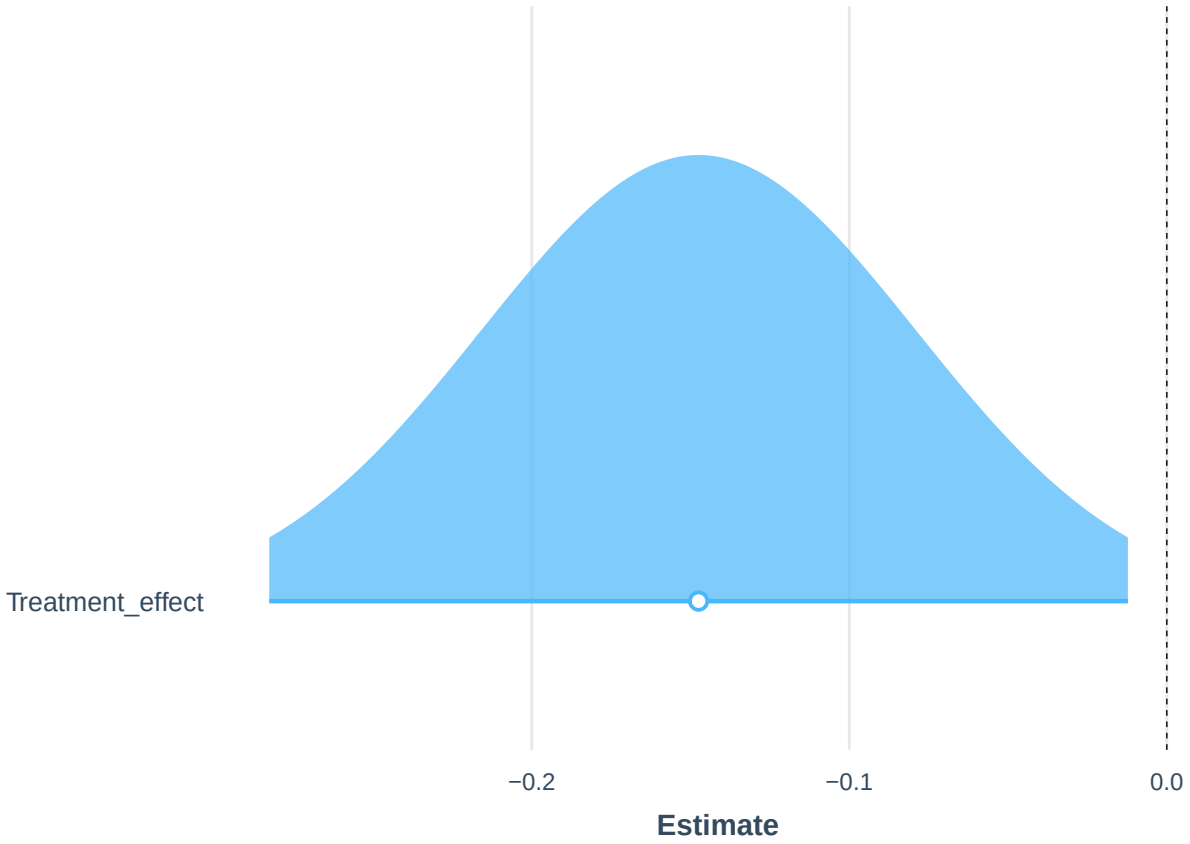
What does the graph suggest about the direction of the effect? The direction of the graph suggest that with the treatment, there is a decrease in the mean share with low perceived theft risk.

(b) To test whether the above difference is compatible with the hypothesis of no effect, i.e. sampling error, we are going to run a regression of the treatment dummy on the outcome variable (using a linear probability model). The estimation equation: Y_i= B_0 + B_1x +B_2e perception_persons_low = Beta_{0} + /beta_{1}*Treatment + error

```
reg1 <- lm(perception_person_low ~ treatment, data=students)
summ(reg1, confint=TRUE)
```

```
## MODEL INFO:
## Observations: 184
## Dependent Variable: perception_person_low
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,182) = 4.63, p = 0.03
## R² = 0.02
## Adj. R² = 0.02
##
## Standard errors: OLS
## ------------------------------------------------------------
##                       Est.    2.5%   97.5%   t val.      p
## ----------------- ------- ------- ------- -------- ------
## (Intercept)           0.40    0.30    0.50     7.96   0.00
## treatment            -0.15   -0.28   -0.01    -2.15   0.03
## ------------------------------------------------------------
```

```
#Plot the 95 percent confidence interval for the estimated treatment effect as follows
treatment_effect <- c("Treatment_effect"="treatment")
plot_summs(reg1, scale = TRUE, coefs = treatment_effect, plot.distributions = TRUE)
```



(d) Did the treatment have a statistically significant effect on the outcome variable? If so, at what level of confidence are you able to reject your null hypothesis (1, 5, 10% confidence level)? Yes the treatment has a statistically significant effect on the outcome variable at a 5% .

(e) Interpret the regression results for the variable treatment. Remember the proper way of reporting an estimated effect from the lecture - and that percentages and percentage points are not the same thing. The chance of viewing the risk of your bike being stolen as low decreases by 15 percentage points when you are reminded of the last time your bike was stolen previously to being asked to asses the risk of your bike being stolen.

f) Is the size of the estimated treatment effect large or small? The estimated treatment effect is -0.15.

Judgement of size of treatment effect: Beta/baseline mean of outcome variable (in control/treatment?) *100% -0.15/0.4*100 = 37.5, this is large.

(g) And in terms of the number of standard deviations of the outcome variable, how does the size of the effect compare to d=0.2 in section VII? R^2 in the regression is 0.02, so the size seems to be appropriate

(h) Those students who never had their bicycle stolen could not really be treated. Does the fact that not all in the treatment group were really treated lead to a bias in the estimated effect of the treatment? No, because the random selection into control/treatment accounted for this (control also has people who have never had their bike stolen).