Assignment 1

Qiaoling Huang (U20421641)

2/1/2020

I collaborated with Salina(Ziqin Ma) for the assignment

```r
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------- tidyverse 1.

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------ tidyverse_conflict
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```r
data = fread("Downloads/AdFX-BA860-SectionA-W20-4004106-rows.csv")
glimpse(data)
```

```
## Observations: 4,004,106
## Variables: 5
## $ Treatment  <int> 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, ...
## $ saw_ads    <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, ...
## $ sales      <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ past_sales <dbl> 0.00, 97.67, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0....
## $ gender     <chr> "female", "male", "male", "female", "male", "male",...
```

Step 1: Convert Gender to 1 and 0

```r
data$gender[data$gender == "female"] <- 1
data$gender[data$gender == "male"] <- 0
glimpse(data)
```

```
## Observations: 4,004,106
## Variables: 5
## $ Treatment  <int> 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, ...
## $ saw_ads    <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, ...
## $ sales      <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ past_sales <dbl> 0.00, 97.67, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0....
## $ gender     <chr> "1", "0", "0", "1", "0", "0", "1", "1", "0", "0", "...
```

```r
data$gender = as.numeric(data$gender)
```

Step 2: Divide data in treatment and control group

```r
treat = data %>% filter(Treatment == 1)
control = data %>% filter(Treatment == 0)

## t-test for user gender
#t.test(treat$gender, control$gender)
```

1: Before analyzing the experiment's results, we want to verify that the experiment properly randomized users. Otherwise, we will not be very confident in our results. To do this, we compare the treatment and control groups by the users' baseline characteristics.

1-a-i: Verify the randomization by user gender

```r
## absolute different and relative different for gender
diff_gender <- mean(treat$gender) - mean(control$gender)
print(diff_gender)
```

```
## [1] -0.0005104218
```

```r
rediff_gender <- diff_gender/mean(control$gender) * 100
print(rediff_gender)
```

```
## [1] -0.1526731
```

```r
## S.E and C.I
diffSE_gender <- sqrt(sd(treat$gender)^2 / length(treat$gender)
           + sd(control$gender)^2 / length(control$gender))
print(diffSE_gender)
```

```
## [1] 0.0005144164
```

```r
ciLow_gender <- diff_gender - 1.96*diffSE_gender
ciHigh_gender <- diff_gender + 1.96*diffSE_gender
print(ciLow_gender)
```

```
## [1] -0.001518678
```

```r
print(ciHigh_gender)
```

```
## [1] 0.0004978343
```

1-a-ii: Report: Absulute different: -0.0005104218 Relative different: -0.1526731% Different of standard error: 0.0005144164 Confidence interval:[-0.001518678, 0.0004978343] Since the confidence interval includes 0, there is no significant different in gender between both group.

1-b-i: Verify the randomization by past sales

```r
## absolute different and relative different for past sales
diff_psales <- mean(treat$past_sales) - mean(control$past_sales)
print(diff_psales)
```

```
## [1] -0.01246596
```

```r
rediffse_psales <- diff_psales/mean(control$past_sales) * 100
print(rediffse_psales)
```

```
## [1] -0.8871942
```

```r
## S.E and C.I
diffSE_psales <- sqrt(sd(treat$past_sales)^2 / length(treat$past_sales)
           + sd(control$past_sales)^2 / length(control$past_sales))
print(diffSE_psales)
```

```
## [1] 0.009706375
```

```r
ciLow_psales <- diff_psales - 1.96*diffSE_psales
ciHigh_psales <- diff_psales + 1.96*diffSE_psales
print(ciLow_psales)
```

```
## [1] -0.03149046
```

```r
print(ciHigh_psales)
```

```
## [1] 0.006558535
```

1-b-ii: Conclusion: Absulute different: -0.01246596 Relative different: -0.8871942% Different of standard error: 0.009706375 Confidence interval:[-0.03149046, 0.006558535] Since the confidence interval includes 0, there is no significant different in past sales between both group.

2: What would you estimate to be the effect of the campaign using an experiment that did not have control ads? Compute the experimental estimate for all users in the experiment: the (average) Intention to Treat estimate. Report the absolute and relative differences, s.e. and c.i..

Answer: It will not be the true effect without control ads, becasue we don't know how many people actually see the ads.

```
## ITT and relative different
ITT <- mean(treat$sales) - mean(control$sales)
print(ITT)
```

```
## [1] -0.001257503
```

```
rediff_ITT <- ITT / mean(control$sales) * 100
print(rediff_ITT)
```

```
## [1] -0.08899183
```

```
## S.E and C.I
diffSE_ITT <- sqrt(sd(treat$sales)^2 / length(treat$sales)
           + sd(control$sales)^2 / length(control$sales))

print(diffSE_ITT)
```

```
## [1] 0.009278338
```

```
ciLow_ITT <- ITT - 1.96*diffSE_ITT
ciHigh_ITT <- ITT + 1.96*diffSE_ITT
print(ciLow_ITT)
```

```
## [1] -0.01944304
```

```
print(ciHigh_ITT)
```

```
## [1] 0.01692804
```

```
##library(lfe)
##summary(felm(sales ~ Treatment, data = data))
```

Conclusion: Absolute different(ITT): -0.001257503 Relative different: -0.08899183% Different of standard error: 0.009278338 Confidence interval:[-0.01944304, 0.01692804] Since the confidence interval includes 0, there is no statistical significant treatment effect for the experience without control ads.

3: This experiment used control ads. Verify that the control ads were deployed the same as the retailer ads by comparing the Treatment and Control groups among the subset of exposed users. 3-a: Verify the equivalence of Treatment exposed and Control exposed users by gender

```
treat_exp <- data %>% filter(Treatment == 1, saw_ads == 1)
control_exp <- data %>% filter(Treatment == 0, saw_ads == 1)
## absolute different
diff_exp_gender <- mean(treat_exp$gender) - mean(control_exp$gender)
print(diff_exp_gender)
```

```
## [1] -0.0004059996
```

```r
## relative different
rediff_exp_gender <- diff_exp_gender / mean(control_exp$gender) * 100
print(rediff_exp_gender)
```

```
## [1] -0.1215191
```

```r
## S.E and C.I
diffSE_exp_gender <- sqrt(sd(treat_exp$gender)^2 / length(treat_exp$gender)
           + sd(control_exp$gender)^2 / length(control_exp$gender))

print(diffSE_exp_gender)
```

```
## [1] 0.0008903802
```

```r
ciLow_exp_gender <- diff_exp_gender - 1.96*diffSE_exp_gender
ciHigh_exp_gender <- diff_exp_gender + 1.96*diffSE_exp_gender
print(ciLow_exp_gender)
```

```
## [1] -0.002151145
```

```r
print(ciHigh_exp_gender)
```

```
## [1] 0.001339146
```

Conclusion: Absulute different: -0.0004059996 Relative different: -0.1215191% Different of standard error: 0.0008903802 Confidence interval:[-0.002151145, 0.001339146] Since the confidence interval includes 0, there is no significant different of exposed users in gender for both group.

3-b: Verify the equivalence of Treatment exposed and Control exposed users by past sales

```r
## absolute different for exposed users in past sales
diff_exp_psales <- mean(treat_exp$past_sales)-mean(control_exp$past_sales)
print(diff_exp_psales)
```

```
## [1] -0.004466455
```

```r
## relative different for exposed users in past sales
rediff_exp_psales <- diff_exp_psales / mean(control_exp$past_sales) * 100
print(rediff_exp_psales)
```

```
## [1] -0.3927394
```

```r
## S.E and C.I for exposed users in past sales
diffSE_exp_psales <- sqrt(sd(treat_exp$past_sales)^2 / length(treat_exp$past_sales)
           + sd(control_exp$past_sales)^2 / length(control_exp$past_sales))
print(diffSE_exp_psales)
```

```
## [1] 0.01162476
```

```r
ciLow_exp_psales <- diff_exp_psales - 1.96*diffSE_exp_psales
ciHigh_exp_psales <- diff_exp_psales + 1.96*diffSE_exp_psales
print(ciLow_exp_psales)
```

```
## [1] -0.02725099
```

```r
print(ciHigh_exp_psales)
```

```
## [1] 0.01831808
```

Conclusion: Absulute different: -0.004466455 Relative different: -0.3927394% Different of standard error: 0.01162476 Confidence interval:[-0.02725099, 0.01831808] Since the confidence interval includes 0, there is no significant different of exposed users in past sales for both group.

4: How does your ad effectiveness estimate change when you make use of the control ads? Compute the experimental estimate for those users who saw ads: the (average) Treatment on Treated (TOT) estimate. Report the absolute and relative differences, s.e. and c.i..

```r
## TOT
TOT = mean(treat_exp$sales) - mean(control_exp$sales)
print(TOT)
```

```
## [1] 0.04050449
```

```r
## Relative different
rediff_TOT <- TOT / mean(control_exp$sales) * 100
print(rediff_TOT)
```

```
## [1] 3.619089
```

```r
## S.E and C.I
diffSE_TOT <- sqrt(sd(treat_exp$sales)^2 / length(treat_exp$sales)
           + sd(control_exp$sales)^2 / length(control_exp$sales))

print(diffSE_TOT)
```

```
## [1] 0.0110392
```

```r
ciLow_TOT <- TOT - 1.96*diffSE_TOT
ciHigh_TOT <- TOT + 1.96*diffSE_TOT
print(ciLow_TOT)
```

```
## [1] 0.01886766
```

```r
print(ciHigh_TOT)
```

```
## [1] 0.06214133
```

```
## summary(felm(sales ~ Treatment, data = data[saw_ads == 1]))
```

Report: TOT for sales Absulute different(TOT): 0.04050449 Relative different: 3.619089% Different of standard error: 0.01162476 Confidence interval:[0.01886766, 0.06214133] Since the confidence interval excludes 0, the treatment effect is statistical significant.

5: What is the total effect of the campaign on sales? (reminder: still include standard error, etc.) 5-a:

```
## Total ITT Lift
Total_ITT <- ITT * nrow(treat)
print(Total_ITT)
```

```
## [1] -3524.765
```

```
## Relative different for total ITT Lift
rediff_Total_ITT <- rediff_ITT
print(rediff_Total_ITT)
```

```
## [1] -0.08899183
```

```
## Total ITT Lift SE
Total_ITT_SE <- diffSE_ITT * nrow(treat)
print(Total_ITT_SE)
```

```
## [1] 26007.06
```

```
## Total ITT Lift C.I
ciLow_Total_ITT <- Total_ITT - 1.96*Total_ITT_SE
ciHigh_Total_ITT <- Total_ITT + 1.96*Total_ITT_SE
print(ciLow_Total_ITT)
```

```
## [1] -54498.6
```

```
print(ciHigh_Total_ITT)
```

```
## [1] 47449.07
```

Report: Absulute different(Total ITT Lift): -3524.765 Relative different: -0.08899183% Different of standard error: 26007.06 Confidence interval:[-54498.6, 47449.07]

5-b:

```
## Total TOT Lift
Total_TOT <- TOT * nrow(treat_exp)
print(Total_TOT)
```

```
## [1] 37911.72
```

```r
## Relative different for total TOT Lift
rediff_Total_TOT <- rediff_TOT
print(rediff_Total_TOT)
```

```
## [1] 3.619089
```

```r
## Total TOT Lift SE
Total_TOT_SE <- diffSE_TOT * nrow(treat_exp)
print(Total_TOT_SE)
```

```
## [1] 10332.56
```

```r
## Total TOT Lift C.I
ciLow_Total_TOT <- Total_TOT - 1.96*Total_TOT_SE
ciHigh_Total_TOT <- Total_TOT + 1.96*Total_TOT_SE
print(ciLow_Total_TOT)
```

```
## [1] 17659.9
```

```r
print(ciHigh_Total_TOT)
```

```
## [1] 58163.54
```

Report: Absulute different: 37911.72 Relative different: 3.619089% Different of standard error: 10332.56 Confidence interval:[17659.9, 58163.54]

5-c: Answer: Based on Question 3, our analysis shows there is no significant different among the subset of exposed users between treatment and control group, meaning that control ads are valid. Since control ads works, we should report TOT estimate from the experiment.

5-d: Answer: Dear Manager, we suggest we should use the TOT estimator since the control ads are valid, our experiment design achieves randonmization for both treatment and control group. And also the ads effect on sales is statistically significant since the confidence interval excludes 0, and the ads effect on sales is 0.04050449, and the total ads effect on sales is 37911.72 in dollars. Therefore, we should launch the ads campaign.

6: What would you estimate to be the effect of the campaign without an experiment? 6-a:

```r
treat_unexp<- treat %>% filter(saw_ads == 0)
## observational absolute different
diff_obs <- mean(treat_exp$sales) - mean(treat_unexp$sales)
print(diff_obs)
```

```
## [1] -0.3784898
```

```r
## Relative different
rediff_obs <- diff_obs/mean(treat_unexp$sales) * 100
print(rediff_obs)
```

```
## [1] -24.60627
```

```
## S.E and C.I
diffSE_obs <- sqrt(sd(treat_exp$sales)^2 / length(treat_exp$sales)
           + sd(treat_unexp$sales)^2 / length(treat_unexp$sales))
print(diffSE_obs)
```

```
## [1] 0.009086299
```

```
ciLow_obs <- diff_obs - 1.96*diffSE_obs
ciHigh_obs <- diff_obs + 1.96*diffSE_obs
print(ciLow_obs)
```

```
## [1] -0.3962989
```

```
print(ciHigh_obs)
```

```
## [1] -0.3606806
```

Report: Absulute different: -0.3784898 Relative different: -24.60627% Different of standard error: 0.009086299 Confidence interval:[-0.3962989, -0.3606806]

6-b: Answer: It could create bias as exposed users may differ unexposed users for the obervantional estimate.

7: Consider gender as a segmentation variable.

7-a:

```
treat_fexp <- treat %>% filter(saw_ads == 1, gender == 1)
control_fexp <- control %>% filter(saw_ads== 1, gender == 1)

## average effect for exposed women (absolute different)
diff_fexp <- mean(treat_fexp$sales) - mean(control_fexp$sales)
print(diff_fexp)
```

```
## [1] 0.1212856
```

```
## Relative different
rediff_fexp <- diff_fexp/mean(control_fexp$sales) * 100
print(rediff_fexp)
```

```
## [1] 7.778076
```

```
## S.E and C.I
diffSE_fexp <- sqrt(sd(treat_fexp$sales)^2 / length(treat_fexp$sales)
           + sd(control_fexp$sales)^2 / length(control_fexp$sales))
print(diffSE_fexp)
```

```
## [1] 0.02459578
```

```r
ciLow_fexp <- diff_fexp - 1.96*diffSE_fexp
ciHigh_fexp <- diff_fexp + 1.96*diffSE_fexp
print(ciLow_fexp)
```

```
## [1] 0.07307788
```

```r
print(ciHigh_fexp)
```

```
## [1] 0.1694933
```

Conclusion: Average effect for exposed women Absulute different: 0.1212856 Relative different: 7.778076% Different of standard error: 0.02459578 Confidence interval:[0.07307788, 0.1694933]

7-b:

```r
treat_mexp <- treat %>% filter(saw_ads == 1, gender == 0)
control_mexp <- control %>% filter(saw_ads== 1, gender == 0)

## average effect for exposed women (absolute different)
diff_mexp <- mean(treat_mexp$sales) - mean(control_mexp$sales)
print(diff_mexp)
```

```
## [1] 0.0004504276
```

```r
## Relative different
rediff_mexp <- diff_mexp/mean(control_mexp$sales) * 100
print(rediff_mexp)
```

```
## [1] 0.05013897
```

```r
## S.E and C.I
diffSE_mexp <- sqrt(sd(treat_mexp$sales)^2 / length(treat_mexp$sales)
           + sd(control_mexp$sales)^2 / length(control_mexp$sales))
print(diffSE_mexp)
```

```
## [1] 0.01103395
```

```r
ciLow_mexp <- diff_mexp - 1.96*diffSE_mexp
ciHigh_mexp <- diff_mexp + 1.96*diffSE_mexp
print(ciLow_mexp)
```

```
## [1] -0.02117611
```

```r
print(ciHigh_mexp)
```

```
## [1] 0.02207697
```

Conclusion: Average effect for exposed men Absulute different: 0.0004504276 Relative different: 0.05013897% Different of standard error: 0.01103395 Confidence interval:[-0.02117611, 0.02207697]

7-c: Answer: We would recommend allocating more budget on female. Because, the confidance interval excludes 0 in female group, but doesn't excludes in man group, meaning that the ads are valid for female group. Moever, the ads effect on sale in female group(0.1212856) is larger than the effect on sale in man(0.0004504276), and the relative different in female(7.778076%) is also larger than the relative different in man group(0.05013897). Meaning that the ads are more effective in female group, therefore, we should allocate more budget in female.