Digital Marketing Analytics Assignment 1

Yixuan Wang

6/9/2021

Online Campaign Evaluation

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

a. Verify the randomization by user gender

i. (10 pts) Use the experimental differences estimator to compare the proportion of women in the Treatment versus Control groups. (Hint: convert the data to numeric using 1=female & 0=male to compute the difference).

```
# Read the data
PS1 <- fread('/Users/wangyixuan/Downloads/PS1-4003988.csv')
PS1$gender_num <- ifelse(PS1$gender=="female", 1, 0)

t.test(PS1[Treatment == 1, gender_num],
       PS1[Treatment == 0, gender_num])
```

```
##
##  Welch Two Sample t-test
##
## data:  PS1[Treatment == 1, gender_num] and PS1[Treatment == 0, gender_num]
## t = -0.69961, df = 2275317, p-value = 0.4842
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0014438006  0.0006842058
## sample estimates:
## mean of x mean of y
## 0.4541049 0.4544847
```

```
# mean difference
gender_mean_diff <- abs(mean(PS1[Treatment == 1, gender_num]) -
                    mean(PS1[Treatment == 0, gender_num]))
cat("Absolute difference in female proportion:", round(gender_mean_diff,3),"\n")
```

```
## Absolute difference in female proportion: 0
```

```r
# relative difference
gender_relative_diff <- abs((mean(PS1[Treatment == 1, gender_num]) -
                             mean(PS1[Treatment == 0, gender_num])))/
                           mean(PS1[Treatment == 0, gender_num])
cat("Relative difference in female proportion:",
    percent(gender_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference in female proportion: 0.084%
```

```r
# SE
gender_se1 <- sd(PS1[Treatment == 1, gender_num])/sqrt(nrow(PS1[Treatment == 1]))
gender_se2 <- sd(PS1[Treatment == 0, gender_num])/sqrt(nrow(PS1[Treatment == 0]))
gender_se_diff <- sqrt((gender_se1)^2 + (gender_se2)^2)
cat("Standard error of difference in female proportion:", round(gender_se_diff, 3),"\n")
```

```
## Standard error of difference in female proportion: 0.001
```

```r
# 95% CI
cat("95% CI of difference in female proportion:[",
    round(gender_mean_diff - 1.96 * gender_se_diff, 3), "," ,
    round(gender_mean_diff + 1.96 * gender_se_diff, 3),"]")
```

```
## 95% CI of difference in female proportion:[ -0.001 , 0.001 ]
```

ii. (5 pts) What do you conclude about the validity of the experimental randomization in terms of gender?

After perform the randomization check on our dataset for the proportion of women, the p-values are all higher than 5% and the differences between treatment and control are small, so most differences are insignificant and we have proper randomization.

b. Verify the randomization by past sales

i. (10 pts) Use the experimental differences estimator to compare the average sales in the 2 weeks before the experiment in the Treatment versus Control groups.

```r
t.test(PS1[Treatment == 1, past_sales],
       PS1[Treatment == 0, past_sales])
```

```
##
##  Welch Two Sample t-test
##
## data:  PS1[Treatment == 1, past_sales] and PS1[Treatment == 0, past_sales]
## t = -0.3029, df = 2275103, p-value = 0.762
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02176521  0.01593835
## sample estimates:
## mean of x mean of y
##  1.471254  1.474167
```

```r
# past sales mean difference
past_mean_diff <- abs(mean(PS1[Treatment == 1, past_sales]) -
                  mean(PS1[Treatment == 0, past_sales]))
cat("Absolute difference in past sales:", round(past_mean_diff,3),"\n")
```

```
## Absolute difference in past sales: 0.003
```

```r
# relative difference
past_relative_diff <- abs((mean(PS1[Treatment == 1, past_sales]) -
                         mean(PS1[Treatment == 0, past_sales])))/
                      mean(PS1[Treatment == 0, past_sales])
cat("Relative difference in past sales:",
    percent(past_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference in past sales: 0.198%
```

```r
# SE
past_se1 <- sd(PS1[Treatment == 1, past_sales])/sqrt(nrow(PS1[Treatment == 1]))
past_se2 <- sd(PS1[Treatment == 0, past_sales])/sqrt(nrow(PS1[Treatment == 0]))
past_se_diff <- sqrt((past_se1)^2 + (past_se2)^2)
cat("Standard error of difference in past sales:", round(past_se_diff, 3),"\n")
```

```
## Standard error of difference in past sales: 0.01
```

```r
# 95% CI
cat("95% CI of difference in past sales:[",
    round(past_mean_diff - 1.96 * past_se_diff, 3), "," ,
    round(past_mean_diff + 1.96 * past_se_diff, 3),"]")
```

```
## 95% CI of difference in past sales:[ -0.016 , 0.022 ]
```

ii. (5 pts) What do you conclude about the validity of the experimental randomization in terms of past sales?

After perform the randomization check on our dataset for past sales variable, the p-values are all higher than 5% and the differences between treatment and control are small, so most differences are insignificant and we have proper randomization.

2. (10 pts) What would be your estimate for 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.

```r
# sales mean difference
sales_mean_diff <- abs(mean(PS1[Treatment == 1, sales]) -
                  mean(PS1[Treatment == 0, sales]))
cat("(average) Intention-to-Treat estimate:", round(sales_mean_diff,3),"\n")
```

```
## (average) Intention-to-Treat estimate: 0.019
```

```r
# relative difference
sales_relative_diff <- abs((mean(PS1[Treatment == 1, sales]) -
                          mean(PS1[Treatment == 0, sales])))/
                        mean(PS1[Treatment == 0, sales])
cat("Relative difference:",
    percent(sales_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference: 1.339%
```

```r
# SE
sales_se1 <- sd(PS1[Treatment == 1, sales])/sqrt(nrow(PS1[Treatment == 1]))
sales_se2 <- sd(PS1[Treatment == 0, sales])/sqrt(nrow(PS1[Treatment == 0]))
sales_se_diff <- sqrt((sales_se1)^2 + (sales_se2)^2)
cat("Standard error:", round(sales_se_diff, 3),"\n")
```

```
## Standard error: 0.009
```

```r
# 95% CI
cat("95% CI:[",
    round(sales_mean_diff - 1.96 * sales_se_diff, 3), "," ,
    round(sales_mean_diff + 1.96 * sales_se_diff, 3),"]")
```

```
## 95% CI:[ 0.002 , 0.037 ]
```

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.

a. (15 pts) Verify the equivalence of Treatment exposed and Control exposed users by gender (repeat both steps in question 1A).

```r
t.test(PS1[Treatment == 1 & saw_ads==1, gender_num],
       PS1[Treatment == 0 & saw_ads==1, gender_num])
```

```
##
##  Welch Two Sample t-test
##
## data:  PS1[Treatment == 1 & saw_ads == 1, gender_num] and PS1[Treatment == 0 & saw_ads == 1, gender_r
## t = -0.79032, df = 1031114, p-value = 0.4293
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002217454  0.000943046
## sample estimates:
## mean of x mean of y
## 0.4544355 0.4550727
```

```r
# mean difference
gender_mean_diff <- abs(mean(PS1[Treatment == 1 & saw_ads==1, gender_num]) -
                    mean(PS1[Treatment == 0 & saw_ads==1, gender_num]))
cat("Absolute difference in female proportion:", round(gender_mean_diff,3),"\n")
```

```
## Absolute difference in female proportion: 0.001
```

```r
# relative difference
gender_relative_diff <- abs((mean(PS1[Treatment == 1 & saw_ads==1, gender_num]) -
                          mean(PS1[Treatment == 0 & saw_ads==1, gender_num])))/
                        mean(PS1[Treatment == 0 & saw_ads==1, gender_num])
cat("Relative difference in female proportion:",
    percent(gender_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference in female proportion: 0.140%
```

```r
# SE
gender_se1 <- sd(PS1[Treatment == 1 & saw_ads==1, gender_num])/
              sqrt(nrow(PS1[Treatment == 1 & saw_ads==1]))
gender_se2 <- sd(PS1[Treatment == 0 & saw_ads==1, gender_num])/
              sqrt(nrow(PS1[Treatment == 0 & saw_ads==1]))
gender_se_diff <- sqrt((gender_se1)^2 + (gender_se2)^2)
cat("Standard error difference in female proportion:", round(gender_se_diff, 3),"\n")
```

```
## Standard error difference in female proportion: 0.001
```

```r
# 95% CI
cat("95% CI difference in female proportion:[",
    round(gender_mean_diff - 1.96 * gender_se_diff, 3), "," ,
    round(gender_mean_diff + 1.96 * gender_se_diff, 3),"]")
```

```
## 95% CI difference in female proportion:[ -0.001 , 0.002 ]
```

b. (15 pts) Verify the equivalence of Treatment exposed and Control exposed users by past sales (repeat both steps in question 1B).

```r
t.test(PS1[Treatment == 1 & saw_ads==1, past_sales],
       PS1[Treatment == 0 & saw_ads==1, past_sales])
```

```
##
##  Welch Two Sample t-test
##
## data:  PS1[Treatment == 1 & saw_ads == 1, past_sales] and PS1[Treatment == 0 & saw_ads == 1, past_sal
## t = -0.61483, df = 1065091, p-value = 0.5387
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02701059  0.01411099
## sample estimates:
## mean of x mean of y
##  1.231168  1.237618
```

```r
# past sales mean difference
past_mean_diff <- abs(mean(PS1[Treatment == 1 & saw_ads==1, past_sales]) -
                  mean(PS1[Treatment == 0 & saw_ads==1, past_sales]))
cat("Absolute difference in past sales:", round(past_mean_diff,3),"\n")
```

```
## Absolute difference in past sales: 0.006
```

```r
# relative difference
past_relative_diff <- abs((mean(PS1[Treatment == 1 & saw_ads==1, past_sales]) -
                             mean(PS1[Treatment == 0 & saw_ads==1, past_sales])))/
                          mean(PS1[Treatment == 0 & saw_ads==1, past_sales])
cat("Relative difference in past sales:",
    percent(past_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference in past sales: 0.521%
```

```r
# SE
past_se1 <- sd(PS1[Treatment == 1 & saw_ads==1, past_sales])/
            sqrt(nrow(PS1[Treatment == 1 & saw_ads==1]))
past_se2 <- sd(PS1[Treatment == 0, past_sales])/
            sqrt(nrow(PS1[Treatment == 0 & saw_ads==1]))
past_se_diff <- sqrt((past_se1)^2 + (past_se2)^2)
cat("Standard error difference in past sales:", round(past_se_diff, 3),"\n")
```

```
## Standard error difference in past sales: 0.013
```

```r
# 95% CI
cat("95% CI difference in past sales:[",
    round(past_mean_diff - 1.96 * past_se_diff, 3), "," ,
    round(past_mean_diff + 1.96 * past_se_diff, 3),"]")
```

```
## 95% CI difference in past sales:[ -0.02 , 0.033 ]
```

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.

```r
# sales mean difference
sales_mean_diff <- abs(mean(PS1[Treatment == 1 & saw_ads==1, sales]) -
                  mean(PS1[Treatment == 0 & saw_ads==1, sales]))
cat("(average) Treatment on Treated (TOT) estimate:", round(sales_mean_diff,3),"\n")
```

```
## (average) Treatment on Treated (TOT) estimate: 0.043
```

```r
# relative difference
sales_relative_diff <- abs((mean(PS1[Treatment == 1 & saw_ads==1, sales]) -
                          mean(PS1[Treatment == 0 & saw_ads==1, sales])))/
                        mean(PS1[Treatment == 0 & saw_ads==1, sales])
cat("Relative difference:",
    percent(sales_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference: 3.613%
```

```r
# SE
sales_se1 <- sd(PS1[Treatment == 1 & saw_ads==1, sales])/
              sqrt(nrow(PS1[Treatment == 1 & saw_ads==1]))
sales_se2 <- sd(PS1[Treatment == 0 & saw_ads==1, sales])/
              sqrt(nrow(PS1[Treatment == 0 & saw_ads==1]))
sales_se_diff <- sqrt((sales_se1)^2 + (sales_se2)^2)
cat("Standard error:", round(sales_se_diff, 3),"\n")
```

```
## Standard error: 0.01
```

```r
# 95% CI
cat("95% CI:[",
    round(sales_mean_diff - 1.96 * sales_se_diff, 3), "," ,
    round(sales_mean_diff + 1.96 * sales_se_diff, 3),"]")
```

```
## 95% CI:[ 0.024 , 0.063 ]
```

5. What is the total effect of the campaign on sales?

a. Compute the total effect using the ITT estimate.

```r
ITT <- abs(mean(PS1[Treatment == 1, sales]) - mean(PS1[Treatment == 0, sales]))
ITT* (nrow(PS1))
```

```
## [1] 78012.68
```

b. Compute the total effect using the TOT estimate.

```r
TOT <- abs(mean(PS1[Treatment == 1 & saw_ads==1, sales]) -
                  mean(PS1[Treatment == 0 & saw_ads==1, sales]))
TOT * (nrow(PS1[saw_ads==1]))
```

```
## [1] 78478.92
```

c. Based on your analysis in question 3, which of the two estimates should you report from this experiment and why?

?? Gender and past sales.

d. Using your preferred estimator, summarize your results for a manager. What are the managerial and statistical implications of your results?

6. What would be your estimate for the effect of the campaign without an experiment? (Hint: you wouldn't have control-group data in this case)

a. Compute the observational estimate.

```r
sales_mean_diff <- abs(mean(PS1[saw_ads==1, sales]) -
                  mean(PS1[saw_ads==0, sales]))
cat("observational estimate:", round(sales_mean_diff,3),"\n")
```

```
## observational estimate: 0.442
```

```r
# relative difference
sales_relative_diff <- abs((mean(PS1[saw_ads==1, sales]) -
                          mean(PS1[saw_ads==0, sales])))/
                        mean(PS1[saw_ads==0, sales])
cat("Relative difference:",
    percent(sales_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference: 26.513%
```

```r
# SE
sales_se1 <- sd(PS1[saw_ads==1, sales])/
              sqrt(nrow(PS1[saw_ads==1]))
sales_se2 <- sd(PS1[saw_ads==0, sales])/
              sqrt(nrow(PS1[saw_ads==0]))
sales_se_diff <- sqrt((sales_se1)^2 + (sales_se2)^2)
cat("Standard error:", round(sales_se_diff, 3),"\n")
```

```
## Standard error: 0.008
```

```r
# 95% CI
cat("95% CI:[",
    round(sales_mean_diff - 1.96 * sales_se_diff, 3), "," ,
    round(sales_mean_diff + 1.96 * sales_se_diff, 3),"]")
```

```
## 95% CI:[ 0.427 , 0.458 ]
```

b. Suppose a manager had not run an experiment and only had the observational estimate. What would they get wrong?

They cannot determine whether the sales are influenced by the recent ad or by control ad.

7. Consider gender as a segmentation variable.

a. Using your preferred estimator from question 5c, what is the average ad effect for women?

```r
# sales mean difference
sales_mean_diff <- abs(mean(PS1[Treatment == 1 & gender_num==1, sales]) -
                  mean(PS1[Treatment == 0 & gender_num==1, sales]))
cat("average estimate for women:", round(sales_mean_diff,3),"\n")
```

```
## average estimate for women: 0.023
```

```r
# relative difference
sales_relative_diff <- (mean(PS1[Treatment == 1 & gender_num==1, sales]) -
                          mean(PS1[Treatment == 0 & gender_num==1, sales]))/
                        mean(PS1[Treatment == 0 & gender_num==1, sales])
cat("Relative difference:",
    percent(sales_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference: 1.194%
```

```r
# SE
sales_se1 <- sd(PS1[Treatment == 1 & gender_num==1, sales])/
              sqrt(nrow(PS1[Treatment == 1 & gender_num==1]))
sales_se2 <- sd(PS1[Treatment == 0 & gender_num==1, sales])/
              sqrt(nrow(PS1[Treatment == 0 & gender_num==1]))
sales_se_diff <- sqrt((sales_se1)^2 + (sales_se2)^2)
cat("Standard error:", round(sales_se_diff, 3),"\n")
```

```
## Standard error: 0.016
```

```r
# 95% CI
cat("95% CI:[",
    round(sales_mean_diff - 1.96 * sales_se_diff, 3), "," ,
    round(sales_mean_diff + 1.96 * sales_se_diff, 3),"]")
```

```
## 95% CI:[ -0.009 , 0.055 ]
```

b. Using your preferred estimator from question 5c, what is the average ad effect for men?

```r
# sales mean difference
sales_mean_diff <- abs(mean(PS1[Treatment == 1 & gender_num==0, sales]) -
                  mean(PS1[Treatment == 0 & gender_num==0, sales]))
cat("average estimate for men:", round(sales_mean_diff,3),"\n")
```

```
## average estimate for men: 0.017
```

```r
# relative difference
sales_relative_diff <- abs((mean(PS1[Treatment == 1 & gender_num==0, sales]) -
                          mean(PS1[Treatment == 0 & gender_num==0, sales])))/
                        mean(PS1[Treatment == 0 & gender_num==0, sales])
cat("Relative difference:",
    percent(sales_relative_diff, accuracy = 0.001),"\n")
```

```
## Relative difference: 1.619%
```

```r
# SE
sales_se1 <- sd(PS1[Treatment == 1 & gender_num==0, sales])/
              sqrt(nrow(PS1[Treatment == 1 & gender_num==0]))
sales_se2 <- sd(PS1[Treatment == 0 & gender_num==0, sales])/
              sqrt(nrow(PS1[Treatment == 0 & gender_num==0]))
sales_se_diff <- sqrt((sales_se1)^2 + (sales_se2)^2)
cat("Standard error:", round(sales_se_diff, 3),"\n")
```

```
## Standard error: 0.009
```

```r
# 95% CI
cat("95% CI:[",
    round(sales_mean_diff - 1.96 * sales_se_diff, 3), "," ,
    round(sales_mean_diff + 1.96 * sales_se_diff, 3),"]")
```

```
## 95% CI:[ -0.001 , 0.035 ]
```

c. Summarize the managerial and statistical implications of your results for a manager who needs to decide how to allocate the ad budget across gender. How will you recommend allocating the budget?

give more budget to women.