Assignment 2_AT_SM

Aditya Tomar, Shubham Midha

2022-10-30

Star Digital, a multimedia video service provider which has a growing focus on online advertising wants to understand their effectiveness of it's digital advertising and the return on investments for each of it's media investments it makes. To understand and answer these questions, they have run an online controlled experiment with the help of an online campaign. The campaign was scheduled to run for six websites with the primary objective of increasing the subscription on package sales.

First it established a randomly assigned treatment and control group that is 90%-10% split of the overall customer base. Star digital, arrived at the split based on factors such as baseline conversion rate, campaign reach, minimum lift expected and other statistical test keeping in mind the opportunity cost. The customers in treatment group are presented to the real promotion as the source of inspiration and control group is shown a charity advertisement in its place to guarantee no overflow between the two companions.

Another variation was in terms of the different websites through which the campaign was run. The different websites had different cost of the ad served, with the first type of websites(1-5) charging $25 per thousand impressions and the other one charging $20 per thousand impressions.

Through this experiment - they aim to answer 3 of the following questions: 1. Is online advertising effective for Star Digital? 2. Is there any frequency effect for advertising on purchase? 3. Which sites should Star Digital advertise on? In particular, should it invest in Site 6 or Sites 1 through 5?

Endogeneity concerns:

```
1. Omitted Variable Bias: There can be many external factors that can impact
the impressions on each website and can be correlated to it.
For example: If the ad is for a show about young adults, chances are that
people of that age group are likely to purchase it or cause more impressions
- response rate among all age groups would not be same.
2. Error in Measurement: Since the only thing to calculate is to record an
impression, we can assume that there is no errors in collecting and
recording the data.
3. Simultaneity: We have no reason or evidence which can say that purchasing
the product leads to more impressions on the websites.
```

Test/Control Selection Bias: Here we have been told that the test/control assignment has been done at random . Also, for our dataset on which we are evaluating our results, a choice based split into purchase = 1 and purchase = 0 was done as well and our sample was picked at random. Overall we can assume that there is no selection bias.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
sd <- read.csv("star_digital.csv")

#loading star digital data

#View(sd)

#creating variable to calculate combined impressions in Website 1, Website 2, Website 3, Website 4, Web
sd$imp_1_5 = sd$imp_1 + sd$imp_2 + sd$imp_3 + sd$imp_4 + sd$imp_5

#creating variable to calculate combined impressions in Website 1, Website 2, Website 3, Website 4, Web

sd$imp_1_6 = sd$imp_1_5 + sd$imp_6
```

```
# Explortatory data analysis

#viewing distributions of target variable (purchase) and manipulation variable (test)

table(sd$purchase)
```

```
##
##     0     1
## 12579 12724
```

```
#Verified choice-based sample output that 50% of people purchased Star Digital, while 50% didn't

table(sd$test)
```

```
##
##     0     1
##  2656 22647
```

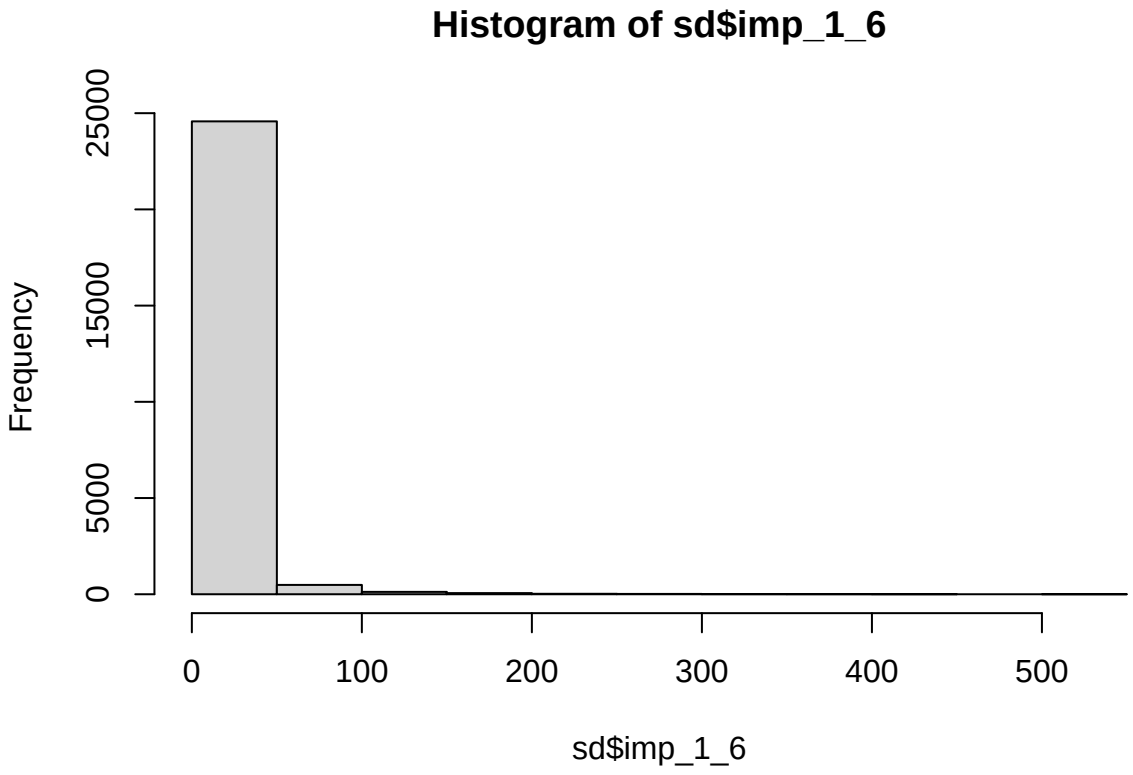
```
#Verified sample constitutes of 10% control and 90% test poeple

table(sd$purchase,sd$test)
```

```
##
##         0     1
##   0  1366 11213
##   1  1290 11434
```

```
#Verified the distribution of test and control across purchasers and non-purchasers is consistent

hist(sd$imp_1_6)
```




```
# No outliers in # of total impressions
```

```
# Checking the randomization efficacy by running a t-test to see if # impressions are similar across te
t.test(imp_1_5~test, data=sd)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_1_5 by test
## t = -0.071371, df = 3268.6, p-value = 0.9431
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8402427  0.7812196
## sample estimates:
## mean in group 0 mean in group 1
##        6.065512        6.095024
```

```
t.test(imp_6~test, data=sd)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_6 by test
## t = 0.43156, df = 2898.4, p-value = 0.6661
```

```
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.3176712  0.4969729
## sample estimates:
## mean in group 0 mean in group 1
##        1.863705        1.774054
```

```
# Verified that #impressions on Website 1-5 & Website 6 are similar across test and control groups (p-v
```

```
# Q1 - Running a linear regression model

summary(lm (purchase ~ test,data=sd))
```

```
##
## Call:
## lm(formula = purchase ~ test, data = sd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5049 -0.5049  0.4951  0.4951  0.5143
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.485693   0.009701  50.064   <2e-16 ***
## test        0.019186   0.010255   1.871   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 25301 degrees of freedom
## Multiple R-squared:  0.0001383,  Adjusted R-squared:  9.882e-05
## F-statistic: 3.501 on 1 and 25301 DF,  p-value: 0.06135
```

```
# This indicates in the sample 48.5% people who are in control group make the purchase, 50.5% people in

# Hence we can not conclude that online advertising is effective (with 95% confidence interval); Howeve
```

```
# For evaluating if the frequency impacts / increases the possibility of purchase, we'll run a linear r

summary(lm (purchase ~ test*imp_1_6,data=sd))
```

```
##
## Call:
## lm(formula = purchase ~ test * imp_1_6, data = sd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89562 -0.47994 -0.05711  0.51280  0.53228
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4651265  0.0101335  45.900  < 2e-16 ***
## test         0.0111885  0.0107209   1.044   0.2967
```

```
## imp_1_6      0.0025937  0.0004131   6.278 3.49e-10 ***
## test:imp_1_6 0.0010362  0.0004408   2.351   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4942 on 25299 degrees of freedom
## Multiple R-squared:  0.02317,    Adjusted R-squared:  0.02306
## F-statistic:   200 on 3 and 25299 DF,  p-value: < 2.2e-16
```

```
# We observe the p-value of the interaction term (0.0188) and coefficient is positive which indicates t
```