Causal-Star Digital

Yu Chun Peng, Chien-Chu Hsu, Chia-Yen Ho

3/17/2022

Introduction and Overview

Star Digital is a large multichannel video service provider with over US$100 million in annual advertising spending. As the advertising market changes, they've gradually increased the share of online advertising spending, hoping to gain the favor of potential customers. At the same time, Star Digital also understood that its return on investment in each ad medium is the key to its spending decisions, so it actively studied the conversion relationship between advertising and sales. Star Digital designed an experiment to test the impact of various online advertising channels on sales using a sample set of over 25,000 online customers. Subjects were randomly assigned to a treatment group, in which they received Star Digital ads, or a control group, in which they received ads from charities on selected websites on ad serving software.

Upon examining the differences between the control and treatment groups, we can see that there is a higher proportion of purchases in the treatment group. However, our findings suggest that the differences cannot be statistically proven, meaning that we cannot be sure whether the Star digital ads are effective in increasing purchases. By studying the relationship between impressions and purchases, we found that impressions have a positive effect on purchases: 1 additional ad impression can lead to a 4% increase in the chance of purchase. Finally, considering which websites to advertise on, we recommend that Star Digital spend its budget on websites 1 through 5 in order to maximize purchase conversion while seeing ads at the lowest cost.

Experimental Design

Star Digital was conducting an experiment on whether their ads were effective on customer purchases using A/B testing. The unit of the experiment was customers. Customers were randomly assigned to two groups, treatment and control. 90% of customers were placed in the treatment group, in which they were shown Star Digital ads. On the other hand, 10% were assigned to the control group, in which they were shown charity ads. The reason that the proportion was disproportionately assigned was to minimize the opportunity cost for potential customers and also to control the cost to display effective ads.

Threats to Causal Inference

1. Selection bias
   The consumers participating in the experiment are not described in our case study, thus it is uncertain whether they are representative. This seems to be a large threat of selection bias in this case because users have a nature intent to purchase.

2. Omitted variable bias
   The consumer demographics are not provided in the case. For example, it is possible that variables such as age is correlated to the subscription decision. Even though we did a randomized experiment and this bias should be controlled, we don't have enough information to conclude this.

3. Simultaneity bias
   In this experiment, the bidirectional effect between the variable and dependent variable is not considered. That is to say, only impressions are considered to influence the purchase decision in the experiment but sometimes purchases may also influence impressions.

4. Measurement error
   How impressions are measured can be problematic to some extent since we don't know if the users actually viewed the ads. For example, many users now install software such as adblockers to hide ads, but if that data was still being recorded, our measurements could be inaccurate.

Exploratory Data Analysis

Before examining causal inference, we began with understanding the data provided by Star Digital. Therefore, we conducted EDA to understand the distribution of data and performed data pre-processing.

Data Description

Below are the columns in the dataset and their definitions.

id: The id of customer
purchase: Customer purchase or not. 0 is no purchase, while 1 is purchase
test: Customer in control or test group. 0 in control group, while 1 in test group
imp_1: The number of ad impressions the customer saw on website 1
imp_2: The number of ad impressions the customer saw on website 2
imp_3: The number of ad impressions the customer saw on website 3
imp_4: The number of ad impressions the customer saw on website 4
imp_5: The number of ad impressions the customer saw on website 5
imp_6: The number of ad impressions the customer saw on website 6

Data Summary

We provided a summary of the dataset to understand the descriptive statistics.

```
summary(df[cols])
```

```
##      imp_1              imp_2             imp_3               imp_4
##  Min.   :  0.0000   Min.   :  0.000   Min.   :  0.00000   Min.   :  0.000
##  1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.00000   1st Qu.:  0.000
##  Median :  0.0000   Median :  0.000   Median :  0.00000   Median :  0.000
##  Mean   :  0.9309   Mean   :  3.428   Mean   :  0.09477   Mean   :  1.589
##  3rd Qu.:  0.0000   3rd Qu.:  2.000   3rd Qu.:  0.00000   3rd Qu.:  0.000
##  Max.   :296.0000   Max.   :373.000   Max.   :148.00000   Max.   :225.000
##      imp_5              imp_6
##  Min.   : 0.00000   Min.   :  0.000
##  1st Qu.: 0.00000   1st Qu.:  0.000
##  Median : 0.00000   Median :  1.000
##  Mean   : 0.04897   Mean   :  1.784
##  3rd Qu.: 0.00000   3rd Qu.:  2.000
##  Max.   :51.00000   Max.   :404.000
```

Check Missing Values

Next, we checked if there is any missing values. After the check, we noticed that there is no missing value in the dataset.

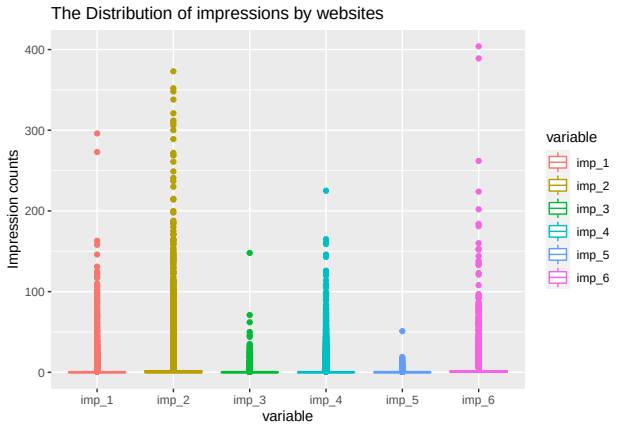
```
sum(is.na(df))
```

```
## [1] 0
```

Check Outliers

Then, we proceeded to check the outliers. To check for outliers regarding impressions, we plotted boxplots by each website to identify the outliers. We noticed that the 25th percentile, the 50th percentile and the 75th percentile is all around 0. Only a few have a value larger than 0. Thus, we would like to clean the data to eliminate extreme outliers. We replaced extreme outliers with the 99.5th percentile values. The reason for this approach is that we still wanted to keep those records instead of eliminating them.

```
new= melt(df,id.vars = "id",measure.vars=c("imp_1","imp_2","imp_3","imp_4","imp_5","imp_6"))
new$variable <- as.character(new$variable)
ggplot(aes(y=value,x=variable, color=variable), data=new) +
geom_boxplot() +
ylab("Impression counts") +
ggtitle("The Distribution of impressions by websites")
```

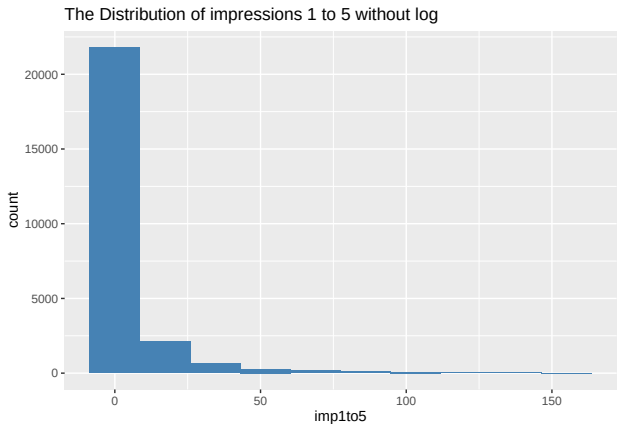


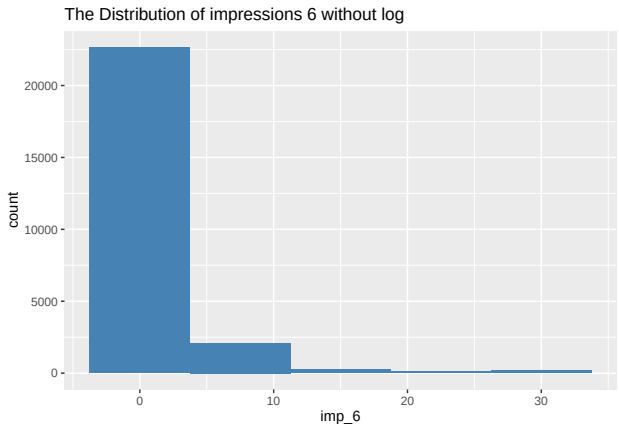
Data Transformation

We plotted a histogram for impressions between websites 1 through website 5 and website 6.

```
ggplot(aes(x=imp1to5), data=df) +
geom_histogram(bins=10,fill="steelblue") +
ggtitle("The Distribution of impressions 1 to 5 without log")

ggplot(aes(x=imp_6), data=df) +
geom_histogram(bins=5,fill="steelblue") +
ggtitle("The Distribution of impressions 6 without log")
```

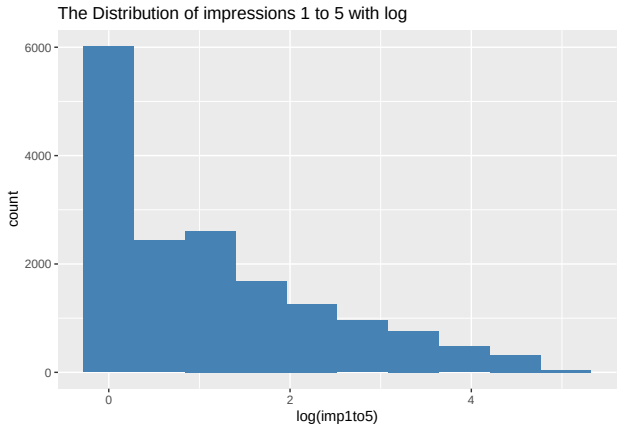


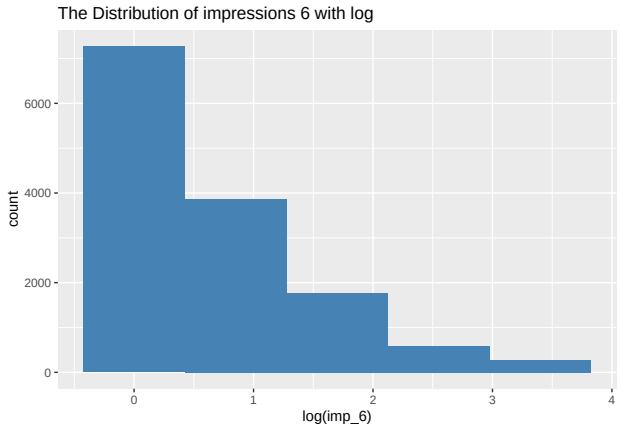


As shown in the plot, plotting histograms for impressions from websites 1-5 and website 6 both show strong negative skew. Therefore, we used log to transform the data and plotted it again. The log transformation provided a better distribution.

```
ggplot(aes(x=log(imp1to5)), data=df) +
geom_histogram(bins=10,fill="steelblue") +
ggtitle("The Distribution of impressions 1 to 5 with log")

ggplot(aes(x=log(imp_6)), data=df) +
geom_histogram(bins=5,fill="steelblue") +
ggtitle("The Distribution of impressions 6 with log")
```





Randomization Test

The first step is to make sure that users are randomized to treatment and control groups, on average there is no difference between these two groups on any characteristics other than treatment. That is, the two groups should be similar in all pre-treatment variables: the number of impressions each user receives.

A t-test is a statistical test that is used to compare the means of two groups. We performed t.test on all the variables individually against the test variables. We wanted to test whether two groups are different from one another.

```r
t.test(imp_all ~ test,data=df, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_all by test
## t = -1.7233, df = 3387.3, p-value = 0.08493
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.06649347  0.06872643
## sample estimates:
## mean in group 0 mean in group 1
##        6.567395        7.066278
```

```r
t.test(imp1to5 ~ test,data=df, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp1to5 by test
## t = -1.5249, df = 3395.4, p-value = 0.1274
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.9534675  0.1191904
## sample estimates:
## mean in group 0 mean in group 1
##        5.057229        5.474367
```

```r
t.test(imp_6 ~ test,data=df, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_6 by test
## t = -1.1438, df = 3311, p-value = 0.2528
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.22186890  0.05837897
## sample estimates:
## mean in group 0 mean in group 1
##        1.510166        1.591911
```

Null hypothesis: There is no difference between the average of impressions between the control and treatment groups.

Alternative hypothesis: There are differences between the average of impressions between the control and treatment groups.

If p-value < 0.05, meaning that the null hypothesis qualifies to be rejected, it indicates that the average impressions between the control and treatment groups are different and are probably not due to chance.

We could see that from the above t-tests, p-values are greater than 0.05, 0.08493 for all impressions, 0.1274 of impressions 1 to 5, 0.2528 for impressions 6. Thus, we failed to reject the null hypothesis: the averages of

impression, sum of impression 1 to 5, and impression 6 are not different between the control and treatment groups. To conclude from the t-tests, we could find that the control and treatment groups have similar impressions to all 6 sites and that users in the control and treatment groups are randomized.

The Effectiveness of Online Advertising for Star Digital

The primary focus of this experiment is to test whether the company's ads have impact on the increase of purchases. The treatment group received the company's ad; on the other hand, the control group receive ads not related to the company. We used t-test to determine whether the treatment actually has an effect on the population of purchase.

```
t.test(purchase~test, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  purchase by test
## t = -1.8713, df = 3309.2, p-value = 0.06139
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.039289257  0.000916332
## sample estimates:
## mean in group 0 mean in group 1
##       0.4856928       0.5048792
```

From the above test result, the mean purchase proportion in the control group is 0.486, and the mean purchase proportion in the treatment group is 0.505, with a difference of approximately 0.019. There is an average 0.019 increase in mean purchase proportion for treatment group.

However, the p-value is 0.06, which is slightly higher than 0.05, we cannot conclude that the mean purchase proportion of treatment groups is significantly higher than that of the control group. That is, we do not know whether online adverting is significantly effective for the company or not. A solution to this is to increase the sample size of the experiment.

Relationship between Ads Impressions and Purchase

We would like to see how much the ad impressions will affect the purchase. So, we used logistic regression to find out how much the change in the number of impressions would result in the change in purchase. Since we only wanted to see the relation of the company's ads and probability of purchase, we filtered out the control group.

```
df_treatment = df %>% filter(test == 1)
m = glm(purchase ~ imp_all, df_treatment , family = "binomial")
summary(m)
```

```
##
## Call:
## glm(formula = purchase ~ imp_all, family = "binomial", data = df_treatment)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.3870  -1.1156   0.2132   1.2229   1.2582
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.228794   0.015837  -14.45   <2e-16 ***
## imp_all      0.040831   0.001618   25.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31393  on 22646  degrees of freedom
## Residual deviance: 30355  on 22645  degrees of freedom
## AIC: 30359
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(summary(m))[2])-1
```

```
## [1] 0.04167643
```

From the above result, we could see that the p-value of imp_all is smaller than 0.05, meaning that the increase in impressions has a significant effect on whether to purchase or not. After transforming the coefficient, we got 0.04, meaning each additional point increase in the company's ad impression increases the odds of purchase by 4%. And it subsequently implied that the more impressions a customer receives, the more likely that a customer would make a purchase.

Choosing between Website 6 or Websites 1 through 5

```
m2 = glm(purchase ~ imp1to5 , df_treatment , family = "binomial")
m3 = glm(purchase ~  imp_6, df_treatment , family = "binomial")

summary(m2)
```

```
##
## Call:
## glm(formula = purchase ~ imp1to5, family = "binomial", data = df_treatment)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4933  -1.1217   0.2053   1.2156   1.2529
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.17583    0.01494  -11.77   <2e-16 ***
## imp1to5      0.04328    0.00178   24.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31393  on 22646  degrees of freedom
## Residual deviance: 30389  on 22645  degrees of freedom
## AIC: 30393
##
## Number of Fisher Scoring iterations: 5
```

```
summary(m3)
```

```
##
## Call:
## glm(formula = purchase ~ imp_6, family = "binomial", data = df_treatment)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6180  -1.1774   0.8291   1.1920   1.1920
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.034302   0.014689  -2.335   0.0195 *
## imp_6        0.034279   0.004055   8.454   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31393  on 22646  degrees of freedom
## Residual deviance: 31317  on 22645  degrees of freedom
## AIC: 31321
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(summary(m2))[2])-1
```

```
## [1] 0.04422638
```

```
exp(coef(summary(m3))[2])-1
```

```
## [1] 0.03487371
```

Each additional point increase in the company's ad impression of websites 1 through 5 increases odds of purchase by 4.4%. Each additional point increase in the company's ad impression of website 6 increases odds of purchase by 3.5%.

```
#cost
cost_imp1to5 <- 25/1000
cost_imp_6 <- 20/1000

#calculate the cost
cost1 <- cost_imp1to5/0.04422638
```

```r
cost2 <- cost_imp_6/0.03487371

cost1
```

```
## [1] 0.5652735
```

```r
cost2
```

```
## [1] 0.5734979
```

As for the cost of advertising on different websites for one thousand impressions, website 1 to 5's cost is \$0.025 per unit impression, while website 6's cost is \$0.02 per unit impression. Then we calculated the cost for 1 unit increase in purchase, which is: 0.565 for website 1 to 5, and 0.573 for website 6.

The cost for websites 1 to 5 is cheaper than website 6. Therefore, we recommended that Star Digital choose to invest money on websites 1 to 5 instead of website 6.

Summary of Recommendation

Based on the experiment design and statistical analysis, we can conclude the following results:

1. We cannot determine whether the advertisement is effective for the company. By testing the difference in purchase between the control and treatment group, we can see that the treatment group, users who see the Star digital ads instead of charity ads, do not have a significantly higher likelihood of purchase. We need further research to examine the difference in purchase between the control and treatment groups.
2. User's exposure to the company's advertisement can significantly increase the chances of purchase. On average, 1 additional ad impression leads to a 4% increase in the chance of purchase.
3. We recommend spending the budget on advertising on sites 1 to 5, as it is more cost-effective to advertise on these sites. On average, to get a 1 unit increase in purchase, the company spent \$0.565 on advertisements on sites 1 through 5. However, the company has to spend \$0.573 on advertisement on site 6 to get the same effects.