Causal - Assignment 2

Abhishikth Peri (5907987), Vaibhav Gakhar

2023-10-07

Business Context

Star Digital is a multichannel video service provider with a significant advertising budget of over $100 million. Starting 2012, they began to gradually increase their spending in online advertising to tap into the consumer market of online media services consumption. However, gauging the true impact of ad-impressions in generating revenues (sales) is a significant hurdle owing to different variables (click, view etc.) associated with the impressions. Understanding this impact and properly tying the two ad-effectiveness metrics (click-based and view-based) is vital for the company's decision-making capability on spending. Hence, Star Digital designed and ran a choice-based advertising experiment on its target consumers in six different websites – aiming to demonstrate the causality between displayed advertisements and generated revenues (sales subscriptions) in the online marketing domain.

Experiment design

In order to measure the incremental effect of advertising on sales and website traffic, Star Digital designed an online advertising campaign - 1) Random assignment – Users were randomly and permanently assigned to either treatment or control groups for the duration of experiment. No user from either group is subjected to advertising of the other. 2) Distinct ad-content - Users from treatment group are exposed to company advertising, while those from control group are exposed to advertisements of a charity organization – as part of the company's normal ad-serving process. Control group served as a baseline against which the impact of Star Digital's advertising could be compared. 3) Control group size – This decision in the experiment was a critical decision influenced by factors like baseline conversion rates, campaign reach, the desired minimum lift for ROI, and the experiment's power. After careful consideration, they chose to allocate 10 percent of users to the control group and 90 percent to the test group - allowing measurement of the impact of display advertising campaign effectively while minimizing costs associated with showing charity ads to potential customers. 4) Campaign Details - The campaign ran on six websites and delivered 170 million impressions to approximately 45 million users over a two-month period in 2012. The primary objective was to increase subscription package sales, but they also aimed to drive traffic to their website to measure brand impact. 5) Company is also interested in conducting a comparitive analysis to identify which of the two buckets amongst Sites 1-5 and Site 6 is more cost-effective for their advertising purposes.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
##
## 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
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.1
```

Load the data set

```
star_digital = read_excel("stardigital.xls")
summary(star_digital)
```

```
##        id             purchase           test           imp_1
##  Min.   :     27   Min.   :0.0000   Min.   :0.000   Min.   :  0.0000
##  1st Qu.: 353881   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:  0.0000
##  Median : 708344   Median :1.0000   Median :1.000   Median :  0.0000
##  Mean   : 708953   Mean   :0.5029   Mean   :0.895   Mean   :  0.9309
##  3rd Qu.:1062738   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:  0.0000
##  Max.   :1413367   Max.   :1.0000   Max.   :1.000   Max.   :296.0000
##      imp_2             imp_3               imp_4             imp_5
##  Min.   :  0.000   Min.   :  0.00000   Min.   :  0.000   Min.   : 0.00000
##  1st Qu.:  0.000   1st Qu.:  0.00000   1st Qu.:  0.000   1st Qu.: 0.00000
##  Median :  0.000   Median :  0.00000   Median :  0.000   Median : 0.00000
##  Mean   :  3.428   Mean   :  0.09477   Mean   :  1.589   Mean   : 0.04897
##  3rd Qu.:  2.000   3rd Qu.:  0.00000   3rd Qu.:  0.000   3rd Qu.: 0.00000
##  Max.   :373.000   Max.   :148.00000   Max.   :225.000   Max.   :51.00000
##      imp_6
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  1.000
##  Mean   :  1.784
##  3rd Qu.:  2.000
##  Max.   :404.000
```

Create a new column 'totalimpressions' to store the cumulative impressions for all websites.

```
star_digital <- star_digital %>% rowwise() %>% mutate(totalimpressions = sum(imp_1,imp_2,imp_3,imp_4,imp
```

Let us check if the how is the purchase data spilt - equally or imbalanced?

```
table(star_digital$purchase)
```

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

Distribution of control group according to purchses made (1) or not (0)

```r
table(star_digital[star_digital$test == 0,]$purchase)
```

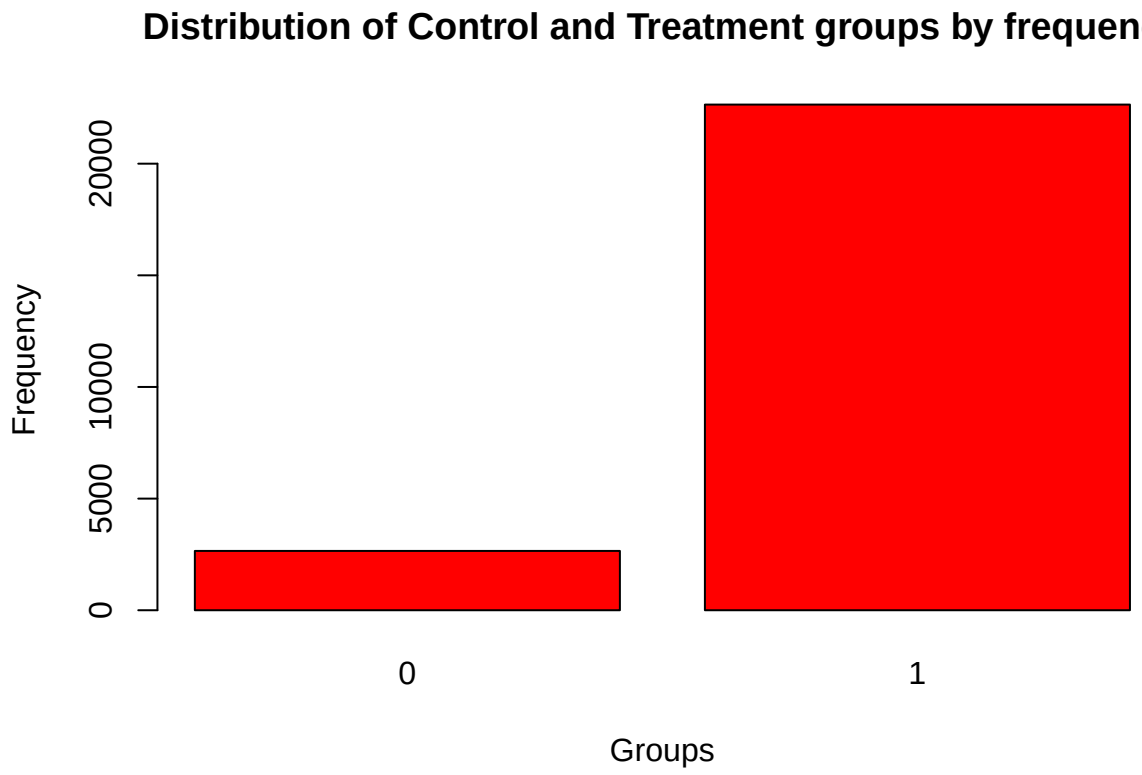
```
##
##    0    1
## 1366 1290
```

Distribution of test group according to purchses made (1) or not (0)

```r
table(star_digital[star_digital$test == 1,]$purchase)
```

```
##
##     0     1
## 11213 11434
```

```r
barplot(table(star_digital$test), main = "Distribution of Control and Treatment groups by frequency", xl
```



Total Count in Control group

```r
cat("Count in Control Group:", sum(star_digital$test == 0), "\n")
```

```
## Count in Control Group: 2656
```

Total Count in treatment group

```r
cat("Count in Treatment Group:", sum(star_digital$test == 1), "\n")
```

```
## Count in Treatment Group: 22647
```

Missing Values Check

```r
missing_data <- colSums(is.na(star_digital))
missing_data
```

```
##               id         purchase             test            imp_1
##                0                0                0                0
##            imp_2            imp_3            imp_4            imp_5
##                0                0                0                0
##            imp_6 totalimpressions
##                0                0
```

As we can see, there are no missing values in the dataset.

#RANDOMIZATION CHECK

Now, we want to check the assumption if the test and control group have a similar treatment (exposure to internet behavior).

```r
t.test(totalimpressions ~ test , data = star_digital)
```

```
##
##  Welch Two Sample t-test
##
## data:  totalimpressions by test
## t = 0.12734, df = 3204.4, p-value = 0.8987
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8658621  0.9861407
## sample estimates:
## mean in group 0 mean in group 1
##        7.929217        7.869078
```

Since the p-value from the hypothesis test (0.8987) $>$ 0.05 (level of significance), we cannot reject null hypothesis, meaning there is no strong evidence to conclude statistically significant difference between treatment group and control group. Hence, Hence, we establish the similarity between treatment and control groups.

#POWER TEST What is the ideal size of both the groups for the experiment? Here, we take power = 0.8, acceptable level of error = 0.1, standard deviation. = 1 and delta (minimum effect size) = 0.1.

```r
power.t.test(sd = 1, power = .8, alternative = "two.sided", sig.level = .05, delta = 0.1,type="two.sampl
```

```
##
##      Two-sample t test power calculation
##
##               n = 1570.737
##           delta = 0.1
```

```
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

In this experiment, the sizes of treatment and control groups are around 23K and 2.6K respectively. From the above power test analysis, the minimum number of users required for each group (treatment and control) to confidently detect a minimum effect size of 0.1 units in the total number of impressions across all websites is 1570. Since our control group of 2.6k (10%) is more than 1570, and our treatment group of 23k (90%) is more than 14130 ((1570*0.9)/0.1 = 14130), we are good with the sample size.

##MAIN ANALYSIS

Q1) Is online advertising effective for Star Digital?

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

```
##
## Call:
## lm(formula = purchase ~ test, data = star_digital)
##
## 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
```

In the above test, we obtain the coefficient of 'test' (1 for treatment and 0 for control) to be 0.019186. This means that when we move from control (test = 0) to treatment (test = 1), there is a likelihood of increase in purchases by 0.019186. However, since the p-value from the hypothesis test (0.0614) > 0.05 (level of significance), there is not enough evidence to conclude the statistical significance of this positive effect on purchases.

Hence, while the test shows that showing Star Digital advertisements (treatment) does have a positive impact on subscription packages purchased, there is not enough evidence to assert the statistical significance of this trend with confidence . ## Q2) Is there a frequency effect of advertising on purchase?

```
summary(lm(purchase ~ test*totalimpressions, data = star_digital))
```

```
##
```

```
## Call:
## lm(formula = purchase ~ test * totalimpressions, data = star_digital)
##
## 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
## totalimpressions        0.0025937  0.0004131   6.278 3.49e-10 ***
## test:totalimpressions 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
```

Here, we look at the effect of treatment, total impressions (frequency) and the interction between them on the consumer purchase of subscription packages made.

1) Exposure to Star Digital advertisements alone increases likelihood of subscription purchases by 0.4651265. However, since the p-value (0.4651265) > 0.05 (level of significance), there is not enough evidence to conclude the statistical significance of this positive effect on purchases.

2) Frequency of ad-impressions alone increases likelihood of subscription purchases by 0.4651265. Additionally, since the p-value (3.49e-10) < 0.05 (level of significance), there is evidence to suggest that increase in frequency of ad-impressions has a positive effect on subscription purchases.

3) The interaction of treatment and frequency of ad-impressions has a positive coefficient of 0.0010362. This means that for combination of both scenarios - when users in treatment group (exposed to Star Digital ads) receive more such advertisements, their likelihood to make purchases increases by 0.0010362. Additionally, since the p-value (0.0188) < 0.05 (level of significance), there is evidence to suggest that increase in frequency of ad-impressions has a positive effect on subscription purchases.

Q3) Which sites should Star Digital advertise on?

Define the cost per thousand impressions for each site

```
s1to5 <- 25 #Cost for Sites 1-5

s6 <- 20 #Cost for Site 6
```

Calculate the total impressions for two buckets - sites 1 to 5 (and) site 6

```
timpressions_s1to5 <- sum(star_digital$imp_1, star_digital$imp_2, star_digital$imp_3, star_digital$imp_
timpressions_s6 <- sum(star_digital$imp_6)
```

Calculate the total costs for two buckets - sites 1 to 5 (and) site 6 Here, we are calculating the cost accrued by Star digital for the total number of impressions garnered through sites 1-5 and site 6 respectively.

```r
tcost_s1to5 <- (timpressions_s1to5 / 1000) * s1to5 #$25 per thousand impressions = $0.025 per impressio
tcost_s6 <- (timpressions_s6 / 1000) * s6 #$20 per thousand impressions = $0.02 per impression
```

Conversions - number of users who purchased subscription package. Here, we are first calculating total number of conversions, then divide total cost of impressions for sites 1-5 and site 6 by total number of conversions to find total cost of conversion for each bucket.

```r
tconversion <- sum(star_digital$purchase)

conversion_s1to5 <- tcost_s1to5 / tconversion
conversion_s6 <- tcost_s6 / tconversion
```

Result - total cost per conversion for sites 1-5

```r
conversion_s1to5
```

```
## [1] 0.3028607
```

Result - total cost per conversion for site 6

```r
conversion_s6
```

```
## [1] 0.0709321
```

It is always preferred to have a lower cost of conversion in a business context. Inferring from the results, site 6 has a lower cost of conversion (0.0709321) compared to sites 1-5 (0.3028607). Hence, Star Digital should consider advertising in Site 6 rather than sites 1-5 to mazimize conversions whilst minimizing advertising spending.

An important point to be considered is that we have only considered impressions to evaluate the advertising spending decision in the above problem. However, in real world. scenarios, ther ecan be multiple other factors such as consumer behavior, external rega=ulatory constraints, geographical considerations etc. which need to be kept in mind while making this decision.

Threats to Causal Inference

1) Omitted variblae bias – Current data only has information regarding impressions; however, we might be missing information. And relations about other variables related to sales.
2) Spillover effect– Indirect influence of treatment users on control user's purchase decision seems unlikely since the sample is choice-based, group allotment does not regard user's individual preference and percentage conversion of users in both groups is very low.
3) Selection bias – Random sampling of the population and choise-based selection of the sample mitigate any selection bias that could be present in the experiment.
4) Measurement Error: We presume that there is no measurement inaccuracy since the only variable recorded at the user level, which is impressions, is straightforward to monitor and track accurately.

KEY FINDINGS

1) The treatment and control groups seems to be statistically similar and well-distributed, paving way for successful inital experimental conditions.

2) Statistical analysis does seem to show positive impact of online advertising of Star Digital on subscription packages, but there is not enough statistical evidence to back this claim with confidence.
3) Treatment (exposure ot Star Digital ads) does seem to show positive impact of online advertising of Star Digital on subscription packages, but there is not enough statistical evidence to back this claim with confidence.
4) Frequency of advertisements seems does seem to show positive impact of online advertising of Star Digital on subscription packages, and there is strong statistical evidence to back this claim.
5) Users in treatment group who are shown more Star digital advertisements seem to show positive likelihood to purchase subscription packages, and there is strong statistical evidence to back this claim.
6) Star Digital has a lower cost per conversion for Site 6 than for sites 1-5.

RECOMMENDATIONS

1) Frequency: Star Digital should prioritize increased frequency of Star Digital ad-impressions on target users, to improve conversions. This can be done through targeted online marketing campaigns, sreamlined advertisement lengths, optimized ad-placement on websites to capture more attention, renewed advertising spending on creating more advertisements etc.
2) Website choice: Star Digital should prioritze allocating more of their advertising budget towards website 6 in order to gain more conversions at lower costs. It is also necessary to monitor the usage and engagement metrics for such platforms to ascertain the further profitability that cn be dervied from Site 6, as well as any potential opportunities to be uncovered from Sites 1-5.
3) Ad-streaming experience: Star Digital should explore obstacles that could occur due to ad-blocking, cookie preventing platforms that can hamper viewing experience of user. Star Digital should also focus on short, relevant and non-intrusive advertisements to increase their ad-impressions whilst engaging their audience with personalization.