Strategic Insights from Star Digital's Advertising Experiment

Yi Hsiang (Royce) Yen

2025-02-23

This analysis investigates the causal impact of Star Digital's online display advertising campaign across six websites, addressing its effectiveness, frequency effects, and optimal site allocation. Using logistic regression on a choice-based sample of 25,303 users, the analysis revealed the following insights:

1. **Online advertising marginally boosts purchase odds by 7.97% (p = 0.0614), though the experiment is severely underpowered (25,303 vs. 3,630,332 required)**.
2. **Increasing ad frequency significantly enhances purchase likelihood, with odds rising 3.3% per impression on Sites 1-5 and 2% on Site 6.**
3. **ROI analysis favors Sites 1-5 (135.2%) over Site 6 (62%), recommending investment in Sites 1-5 for higher returns despite elevated costs.**

These insights guide Star Digital toward optimizing its advertising strategy for maximum effectiveness and profitability.

```
# Load Library & Import Data Set as Data
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
```

```
library(readxl)
data <- read_excel("M347SS-XLS-ENG (1).xls", sheet = 'star.csv')
```

1. Overview of setting

Star Digital designed an online display advertising campaign run on six websites with the primary objective of increasing subscription package sales. In this experiment:

- The unit of analysis is a user/consumer.

- The definition of treatment is exposure to Star Digital's advertising campaign. The test group received advertisements for Star Digital, while the control group received advertisements for a charity organization.
- The experimental design incorporates random assignment to ensure reliable targeting. Before an ad was served, each user was randomly assigned to either the test or control group. The ad-serving software then checked this assignment to determine whether to show the Star Digital ad or the charity ad. A user assigned to the control group was never exposed to a Star Digital ad from the campaign, and vice versa. We can reliably target a unit with treatment. This indicates that we don't have clear evidence to conclude the SUTVA assumption is violated, meaning the treatment of subjects in the treatment group does not affect the treatment of subjects in the control group.
- The data was from a choice-based sample. In this data set, 50 percent consisted of people who had chosen to purchase the subscription package of Star Digital, while the remaining 50 percent consisted of those who had not purchased the package. However, whether the person belonged to the control group or test group was random in this sample. While group assignment was random within the sample, the initial selection based on purchase behavior could introduce selection bias.
- Users were selected based on their purchase decision. (Whether they purchased the subscription package or not.)
- Prior to the experiment, users had not yet been served any campaign ads.
- During the course of the campaign, it delivered 170 million impressions to about 45 million users over a period of two months in 2012.

The key questions to be addressed are as follows:

1. Is online advertising effective for Star Digital?

2. Is there a frequency effect of advertising on purchase? In particular, the question is whether increasing the frequency of advertising increases the probability of purchase?

3. Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in Site 6 or in Sites 1 through 5?

2. Hypothesis and testing

2-1. Do we have enough sample size to identify any statistically significant differences?

The first thing to check is whether our sample size is large enough to detect a statistically significant difference. I will do this with a power analysis.

According to the description, the conversion rate for this campaign was extremely low—only 0.153% of consumers, including both the control and test groups, made a purchase. To estimate the minimum effective sample size, I will assume the ad does have an effect, and assign the desired conversion rate for the control group (p1) as 0.00153, and for the treatment group (p2), I'll assume it to be slightly higher, at 0.001683 This means I am estimating the **minimum sample size required to detect at least a 10% increase in purchase proportion due to the treatment, with at least 80% confidence.**

```
# Check how many records do we have for each group
data %>% group_by(test) %>% summarize(n())
```

```
## # A tibble: 2 x 2
##    test `n()`
```

```
##   <dbl> <int>
## 1     0  2656
## 2     1 22647
```

```r
# Perform power analysis
power.prop.test(p1 = 0.00153, p2 = 0.001683, power = 0.8,
                sig.level = 0.05, alternative = "two.sided")
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 1075565
##              p1 = 0.00153
##              p2 = 0.001683
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

According to the power analysis, this experiment is severely underpowered, meaning we may falsely find that there is no effect. Expected minimum sample size for each group should be 1075565 users, while we only have 2656 users for our control group and 22647 for our treatment group. Therefore, for the following hypothesis tests, when encountering a large p-value, we need to take this into consideration.

2-2. Randomization Check

Next, we have to perform the following randomization checks:

1. **Distribution of ad impressions across websites between the treatment and control group**. This checks whether the only difference between the treatment and control group is the displayed ad (Star Digital ad or the charity ad).
2. **Distribution of ad impressions across websites between the purchasers vs non-purchasers**. This is to ensure that the relationship between purchase behavior and ad exposure is not obscured by differences in exposure levels, allowing for a more accurate understanding of the causal impact of advertising on purchase behavior.

I'll do this with t tests.

```r
# Ad impressions between treatment & control group
t.test(imp_1 ~ test, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_1 by test
## t = -3.905, df = 3574.1, p-value = 9.596e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.5961772 -0.1976257
## sample estimates:
## mean in group 0 mean in group 1
##       0.5756777       0.9725791
```

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

```
##
##  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
```

According to the results, the relatively large p-value (0.6661) suggests that there is no strong evidence of a difference in ad impression distribution for website 6 between the treatment and control groups. **However, choice-based sampling does not guarantee that ad impressions are randomly and evenly distributed across groups.** This could introduce an imbalance in exposure, which might confound the estimated treatment effect.

On the other hand, **ad impressions on some websites (e.g., imp_1) do show a significant difference between groups, suggesting that randomization might not have been fully effective.** As a result, the observed effect of the treatment could be influenced by differences in exposure rather than the ad itself.

Given these results, **I acknowledge the potential impact of choice-based sampling but proceed with the analysis under the assumption that any residual imbalance does not meaningfully affect the overall conclusions. Additional robustness checks may be necessary to validate this assumption.**

```r
# Ad impressions between purchase & non-purchasers
t.test(imp_1 ~ purchase, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_1 by purchase
## t = -10.104, df = 22868, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8505461 -0.5741565
## sample estimates:
## mean in group 0 mean in group 1
##       0.5727005       1.2850519
```

```r
t.test(imp_6 ~ purchase, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_6 by purchase
## t = -6.9498, df = 19654, p-value = 3.772e-12
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.7821661 -0.4380293
## sample estimates:
## mean in group 0 mean in group 1
##        1.476667        2.086765
```

Consistent with the explanation for ad impressions between treatment and control groups, the t-tests for Ad impressions between purchase & non-purchasers shows similar pattern. **I acknowledge the potential impact of choice-based sampling but proceed with the analysis under the assumption that any residual imbalance does not meaningfully affect the overall conclusions. Additional robustness checks may be necessary to validate this assumption.**

2-3. Is online advertising effective for Star Digital

In our case, the target variable(purchase) is a binary variable, while the input/independent variables are interval/numerical. Therefore, I'll test this out with a logistic regression model.

```
model = glm(purchase ~ test, data = data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = purchase ~ test, family = binomial, data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.05724    0.03882  -1.474   0.1404
## test         0.07676    0.04104   1.871   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 35073  on 25301  degrees of freedom
## AIC: 35077
##
## Number of Fisher Scoring iterations: 3
```

```
exp(0.07676)
```

```
## [1] 1.079783
```

If we follow the common threshold 0.05, we might conclude that online advertising is not effective for Star Digital. However, as mentioned earlier, **according to the power analysis, this experiment is severely underpowered, meaning we may falsely find that there is no effect.** As a result, if we follow a more loosen threshold of 0.1, **according to the logistic regression model, the p-value is 0.0614<0.1, meaning there is a relationship between online advertising and purchase.** The positive coefficient 0.07676 indicates that **assuming all other factors remain constant, the odds of purchasing increase by 7.97% when online advertising is implemented (compared to when it is not implemented).**

2-4. Is there a frequency effect of advertising on purchase? In particular, the question is whether increasing the frequency of advertising increases the probability of purchase?

To examine whether increasing advertising frequency increases the probability of purchase, I divided the analysis into three groups:

1. Group 1: Sites 1-5
2. Group 2: Site 6
3. Group 3: All sites combined

This classification is based on Star Digital's control over ad placements. By separating the sites this way, I aim to investigate **whether ad frequency effects vary across different advertising environments.** However, it is important to ensure that within each group, there is sufficient variation in the number of ad impressions received by users.

```
# Define impression groups and check variation with stats and visuals
data <- data %>% mutate(group1 = imp_1 + imp_2 + imp_3 + imp_4 + imp_5,
                        group3 = imp_1 + imp_2 + imp_3 + imp_4 + imp_5 + imp_6)
# Function to compute and print summary stats
print_summary_stats <- function(data, varname, group_label) {
  var <- data[[varname]]
  stats <- summary(var)
  sd_val <- round(sd(var), 2)
  prop_nonzero <- round(mean(var > 0), 2)
  cat(group_label, ":\n")
  cat("Mean =", round(mean(var), 2), ", SD =", sd_val, "\n")
  cat("Min =", stats[1], ", Q1 =", stats[2], ", Median =", stats[3],
      ", Q3 =", stats[5], ", Max =", stats[6], "\n")
  cat("Proportion > 0 =", prop_nonzero, "\n\n")}
# Check variation in impression counts
cat("Variation in Impression Counts:\n")
```

```
## Variation in Impression Counts:
```

```
print_summary_stats(data, "group1", "Sites 1-5 (group1)")
```

```
## Sites 1-5 (group1) :
## Mean = 6.09 , SD = 19.68
## Min = 0 , Q1 = 0 , Median = 1 , Q3 = 4 , Max = 397
## Proportion > 0 = 0.65
```

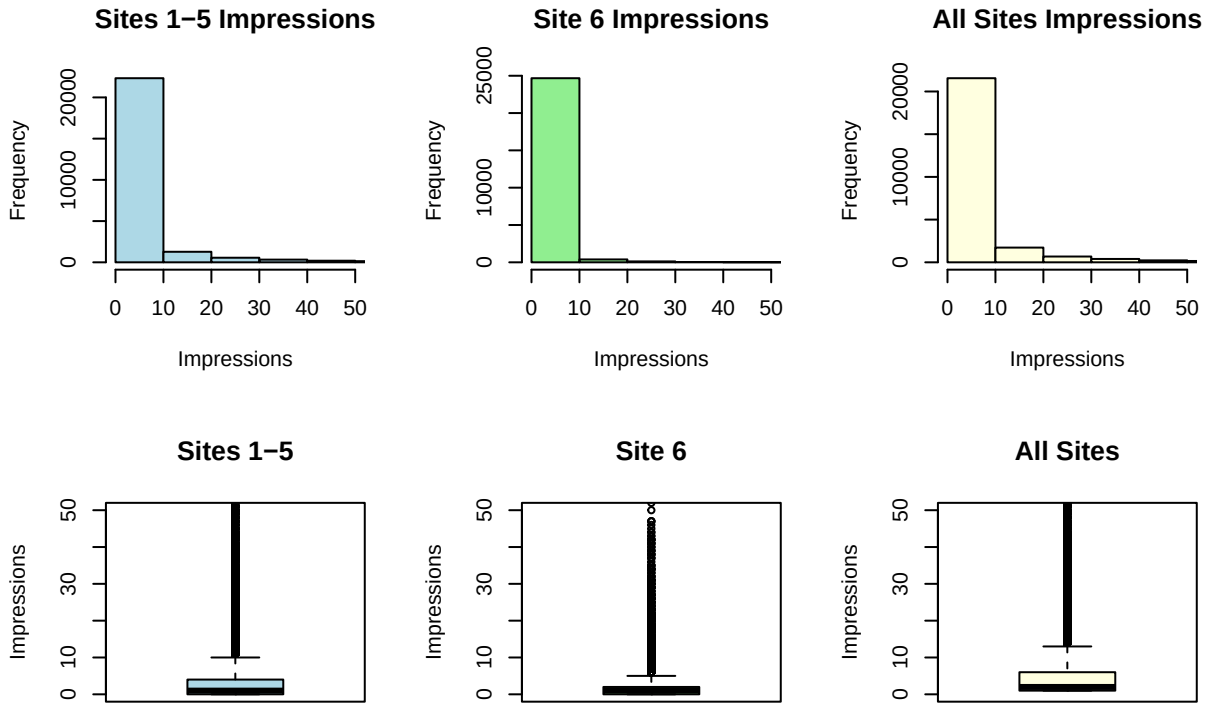
```
print_summary_stats(data, "imp_6", "Site 6 (imp_6)")
```

```
## Site 6 (imp_6) :
## Mean = 1.78 , SD = 7.01
## Min = 0 , Q1 = 0 , Median = 1 , Q3 = 2 , Max = 404
## Proportion > 0 = 0.54
```

```r
print_summary_stats(data, "group3", "All Sites (group3)")
```

```
## All Sites (group3) :
## Mean = 7.88 , SD = 21.56
## Min = 1 , Q1 = 1 , Median = 2 , Q3 = 6 , Max = 521
## Proportion > 0 = 1
```

```r
# Visualizations
par(mfrow = c(2, 3))
# Histograms (capped at 50, adjust if max > 50)
hist(data$group1, breaks = 50, xlim = c(0, 50), main = "Sites 1-5 Impressions",
     xlab = "Impressions", col = "lightblue", border = "black")
hist(data$imp_6, breaks = 50, xlim = c(0, 50), main = "Site 6 Impressions",
     xlab = "Impressions", col = "lightgreen", border = "black")
hist(data$group3, breaks = 50, xlim = c(0, 50), main = "All Sites Impressions",
     xlab = "Impressions", col = "lightyellow", border = "black")
# Box plots (capped at 50, adjust if needed)
boxplot(data$group1, ylim = c(0, 50), main = "Sites 1-5", ylab = "Impressions", col = "lightblue")
boxplot(data$imp_6, ylim = c(0, 50), main = "Site 6", ylab = "Impressions", col = "lightgreen")
boxplot(data$group3, ylim = c(0, 50), main = "All Sites", ylab = "Impressions", col = "lightyellow")
```




```r
par(mfrow = c(1, 1))
cat("Interpretation: Sufficient variation if SD > 0, median > 0, and >50% non-zero impressions.\n")
```

```
## Interpretation: Sufficient variation if SD > 0, median > 0, and >50% non-zero impressions.
```

The observed distribution of impressions, with a high proportion of zeros (35% for Sites 1-5, 46% for Site 6) and extreme outliers (max 397, 404, 521), is most likely due to the selection method of this sample, which oversampled purchasers (50% in sample vs. 0.153% in population) who may have higher impression counts. For the remaining analysis, I will assume there is sufficient variation based on the non-zero SDs (19.68, 7.01, 21.56), medians (1, 1, 2), and proportions > 0 (65%, 54%, 100%), and proceed by ignoring the potential impact of this selection bias on the frequency effect and ROI calculations, noting that this may overestimate the true effect sizes.

Next, I'll move on to examine whether increasing advertising frequency increases the probability of purchase.

```
data <- data%>%mutate(group1 = imp_1 + imp_2 + imp_3 + imp_4 + imp_5,
                      group3 = imp_1 + imp_2 + imp_3 + imp_4 + imp_5 + imp_6)
group1 = glm(purchase ~ group1, data = data, family = binomial)
group2 = glm(purchase ~ imp_6, data = data, family = binomial)
group3 = glm(purchase ~ group3, data = data, family = binomial)
summary(group1)
```

```
##
## Call:
## glm(formula = purchase ~ group1, family = binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.145722   0.013946  -10.45   <2e-16 ***
## group1       0.032438   0.001461   22.21   <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: 35077  on 25302  degrees of freedom
## Residual deviance: 34218  on 25301  degrees of freedom
## AIC: 34222
##
## Number of Fisher Scoring iterations: 5
```

```
summary(group2)
```

```
##
## Call:
## glm(formula = purchase ~ imp_6, family = binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.022103   0.013433  -1.645   0.0999 .
## imp_6        0.019834   0.002927   6.776 1.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35077  on 25302  degrees of freedom
## Residual deviance: 35014  on 25301  degrees of freedom
```

```
## AIC: 35018
##
## Number of Fisher Scoring iterations: 4
```

```
summary(group3)
```

```
##
## Call:
## glm(formula = purchase ~ group3, family = binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.179392   0.014583  -12.30   <2e-16 ***
## group3       0.029201   0.001294   22.56   <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: 35077  on 25302  degrees of freedom
## Residual deviance: 34212  on 25301  degrees of freedom
## AIC: 34216
##
## Number of Fisher Scoring iterations: 5
```

```
exp(0.032438)
```

```
## [1] 1.03297
```

```
exp(0.019834)
```

```
## [1] 1.020032
```

```
exp(0.029201)
```

```
## [1] 1.029632
```

According to the results of the logistic regression models, **for all three groups, the p-value is smaller than 0.05, indicating that there is a statistically significant frequency effect of advertising on the odds of purchasing.**

- For **sites 1 through 5** (Group 1), assuming all other factors remain constant, for each additional impression, **the odds of purchasing increase by 3.297%.**
- For **site 6** (Group 2), assuming all other factors remain constant, for each additional impression, **the odds of purchasing increase by 2%.**
- For the combined data across all websites (Group 3), assuming all other factors remain constant, for each additional impression, **the odds of purchasing increase by 2.92%.**

2-5. Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in Site 6 or in Sites 1 through 5?

Costs are 25 USD per 1,000 impressions (Sites 1-5) and 20 USD (Site 6), with 1,200 USD revenue per purchase. With this information, we can estimate the ROI for Site 6 vs Sites 1 through 5 with the following steps:

1. Establish baseline purchase probability
2. Adjust the conversions variation for site 1 through 5 and site 6
3. Calculate the expected purchase volume per 1000 impressions.
4. Calculate the estimated revenue & ROI

```
# 1. Establish Baseline Probability (P0)
P0 <- 0.00153 # Population baseline from case study: 0.153%
# 2. Recalibrate betas
# Original betas from biased sample (0.032438 for Sites 1-5, 0.019834 for Site 6)
# Ideally, weight sample to match 0.153%, but without weights, adjust with population baseline
beta_group1 <- 0.032438  # Sites 1-5 from group1 model
beta_imp6 <- 0.019834    # Site 6 from group2 model
# Compute beta0 for population P0 = 0.00153
beta0 <- log(P0 / (1 - P0))  # roughly equals to -6.481
# 3. Estimate Delta P per Impression
# Sites 1-5
P1_group1 <- exp(beta0 + beta_group1) / (1 + exp(beta0 + beta_group1))  # roughly equals to 0.001579
delta_P_group1 <- P1_group1 - P0  # roughly equals to 0.000049
# Site 6
P1_imp6 <- exp(beta0 + beta_imp6) / (1 + exp(beta0 + beta_imp6))  # roughly equals to 0.001557
delta_P_imp6 <- P1_imp6 - P0  # roughly equals to 0.000027
# 4. Scale to 1,000 Impressions
purchases_group1 <- 1000 * delta_P_group1  # roughly equals to 0.049 purchases
purchases_imp6 <- 1000 * delta_P_imp6      # roughly equals to 0.027 purchases
# 5. Calculate Revenue and ROI
# Sites 1-5
revenue_group1 <- purchases_group1 * 1200  # roughly equals to 58.8
cost_group1 <- 25
roi_group1 <- (revenue_group1 - cost_group1) / cost_group1  # roughly equals to 1.352 (135.2%)
# Site 6
revenue_imp6 <- purchases_imp6 * 1200  # roughly equals to 32.4
cost_imp6 <- 20
roi_imp6 <- (revenue_imp6 - cost_imp6) / cost_imp6  # roughly equals to 0.62 (62%)
```

Using logistic regression (population baseline 0.153%):

- Sites 1-5: Each impression increases probability by 0.0049%, yielding 0.049 purchases per 1,000 impressions, revenue = $58.80, ROI = 135.2%.
- Site 6: Each impression increases probability by 0.0027%, yielding 0.027 purchases per 1,000 impressions, revenue = $32.40, ROI = 62%.

Note that these ROIs may be slightly inflated due to choice-based sampling; true values could be lower with sample weighting. Nevertheless, according to the ROI, the recommendation is to prioritize Sites 1-5 for higher ROI (135.2% vs. 62%).