Causal Analysis Project- A/B Testing on App's Booking Page

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

4/7/2022

1. Problem Statement

This is a Flyber app that provides an on-demand flying-taxi service in New York City. As a data analytics team, we're dedicated to optimizing the user experience for the Flyber app to enhance the conversion rate from every step of the user journey. The statistics showed the conversion rate of the app's booking page was comparatively lower than the competitors. Therefore, the goal of the test was to see how people engage with the app booking page and which variations generated more bookings.

2. Data Description

The dataset used in our A/B testing analysis was event-level data from a multivariate test for the Flyber app. It is from a Github page. There are approximately 80,000 users and 8 variables in the dataset, namely -

1. **User_uuid**: encrypted user id
2. **Experiment_group**: Control group is denoted as "control" while treatment groups are denoted as "experiment_1", "experiment_2"

- Control group users will be navigated to the original version of the app, with the words "Tip included", and the button "Book Flight" to further proceed
- Treatment group users are navigated to two different app versions -

    a. With "Tip included", "Fly Now" to book
    b. Without "Tip included", "Book Flight" to book

3. **Event_uuid**: Trip booking event id
4. **Event_time**: Time when user triggered a trip
5. **Age**: Four age groups including 18-29, 30-39, 40-49 and 50+
6. **Session_uuid**: Session id
7. **User_neighborhood**: Neighborhoods in New York City such as Brooklyn, Manhattan etc
8. **Event_type**: Last event the user engaged, four steps included "Open app", "Enter number of riders", "Search" and "Begin ride"

3. Experiment Hypothesis

Three different versions of the app were tested out for the highest conversion probability. The user funnel contains four steps - "Open app", "Enter number of riders", "Search", and "Begin ride".

Experiment was conducted on the booking page of the app, which is also the last step of the user journal and the most critical customer experience. It's essential to test and experiment with the funnel to seek improvements that can raise conversion or improve the customer experience.

We hypothesized that a redesign of the text on the booking button would make it more personal and highlight the benefits of the experience would lead to more conversions. Another hypothesis was that removing the text 'tips included' on that page would simplify the page and make it easier for users to navigate.

We tested 2 different versions of the app against the original and wanted to know which version of the app page would bring a higher conversion probability. We chose conversion probability as it is a metric that can affect the high-level business metric (e.g. revenue) and would likely be affected by the proposed changes. The conversion rate is defined as the number of booked rides divided by the number of users who enter the booking page. In addition, we wanted to focus on the conversion probability between the stage "Open", and "Search". This is the early phase that would most likely start engaging our customers.

Hypothesis 1: Redesign the text on the booking button would lead to more conversions (Control vs experiment 1)

Hypothesis 2: Remove the "tip included" text would lead to more conversions (Control vs experiment 2)

4. Threats to Causal Inference

1. Selection bias Research is usually conducted in a subset of the population, either out of necessity or convenience. Selection bias can result when the selected portion of the population differs from the total population in terms of exposure and outcome of interest. There may be a huge threat of selection bias in our case since users could share the same preference or habit of riding.

2. Omitted variable bias We're suspecting that the dataset doesn't include some of the considerable variables. For example, it's likely the user's decision would be influenced by their income. 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 the booking button design is considered to influence the users' decisions in the experiment. Therefore, there is no simultaneity bias in this case.

4. Measurement error The measurement of the experiment can be problematic to some extent since we're not sure if the users really notice the change in button design and text. It's possible those users are too used to the original user interface so they didn't pay attention to the new design. Another possibility is that their friends place the order for them. So, of course, this would affect our experiment given the inaccurate data.

5. Data Cleaning

Before examining causal inference, we began with the preparation of data by data cleaning.

```
data$event_day = as.Date(data$event_time, format = "%m/%d/%y %H:%M")
data$open = ifelse(data$event_type == 'open', 1, 0)
data$of_users = ifelse(data$event_type == '#_of_users', 1, 0)
```

```
data$search = ifelse(data$event_type == 'search', 1, 0)
data$begin_ride = ifelse(data$event_type == 'begin_ride', 1, 0)

#Number of users in each funnel looks like below:
table(data$event_type) %>%
  as.data.frame() %>%
  arrange(desc(Freq))  %>%
  setNames(c("Stages", "Number of users"))
```

```
##       Stages Number of users
## 1       open          114853
## 2 #_of_users           48224
## 3     search           23134
## 4 begin_ride             345
```

The number of users in each stage is shown in the chart. We wanted to focus on the conversion probability between the stage "Open", and "Search" as this is the early phase that would most likely start engaging our customers.

6. Explore the data with Visualization

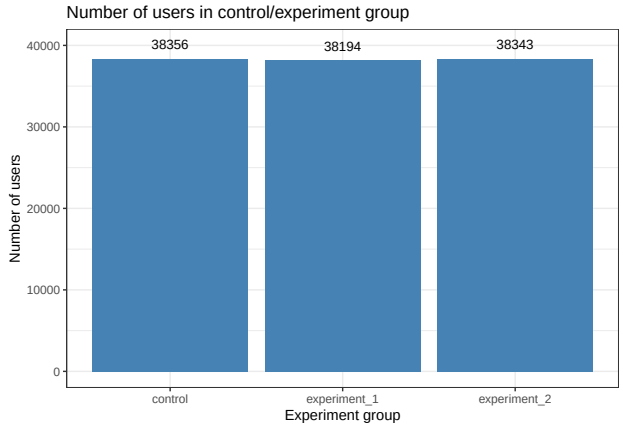
Next, we proceed to understand the data by conducting EDA to understand the distribution of data.

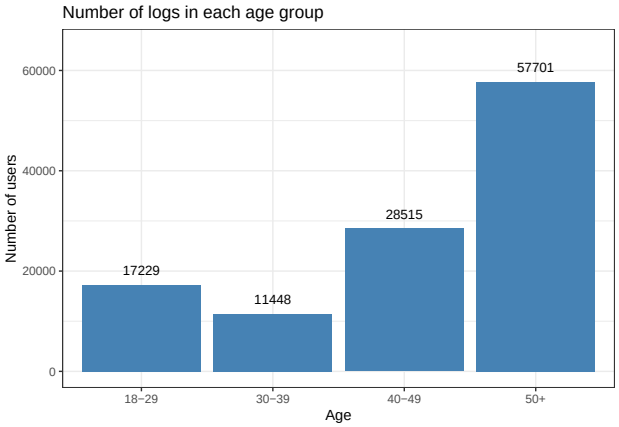
The Distribution of Variables

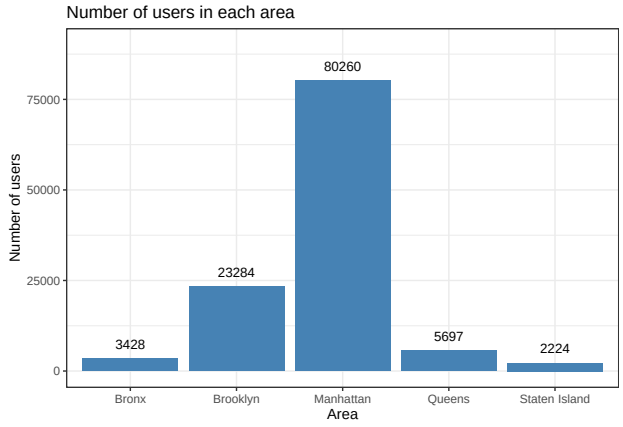
We plotted histograms for the number of users in different groups, the distribution of age and the distribution of neighborhood.

```
df_experiment_group = table(df2$experiment_group)%>%
  as.data.frame()

ggplot(data=df_experiment_group, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue") +
  labs(title="Number of users in control/experiment group", x = "Experiment group", y = "Number of users
  ylim(0, 40000) +
  geom_text(aes(label=Freq), vjust=-1, color="black", size=3.5) +
  theme_bw()

# age
df_age = table(df2$age)%>%
  as.data.frame()

ggplot(data=df_age, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue") +
  labs(title="Number of logs in each age group", x = "Age", y = "Number of users") +
  ylim(0, 65000) +
  geom_text(aes(label=Freq), vjust=-1, color="black", size=3.5) +
  theme_bw()

# neighborhood
```

```
df_neighborhood = table(df2$user_neighborhood)%>%
  as.data.frame()%>%
  arrange(desc(Freq))

ggplot(data=df_neighborhood, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue") +
  labs(title="Number of users in each area", x = "Area", y = "Number of users") +
  ylim(0, 90000) +
  geom_text(aes(label=Freq), vjust=-1, color="black", size=3.5) +
  theme_bw()
```







Comparisons of Different Experiments

We plotted line charts to compare the users in the 'Open' and 'Search' phases among the control and treatment groups. We also plotted the conversion rate.

```
df1 = data %>% group_by(experiment_group, event_day, event_type) %>% summarise(cnt = n())
```

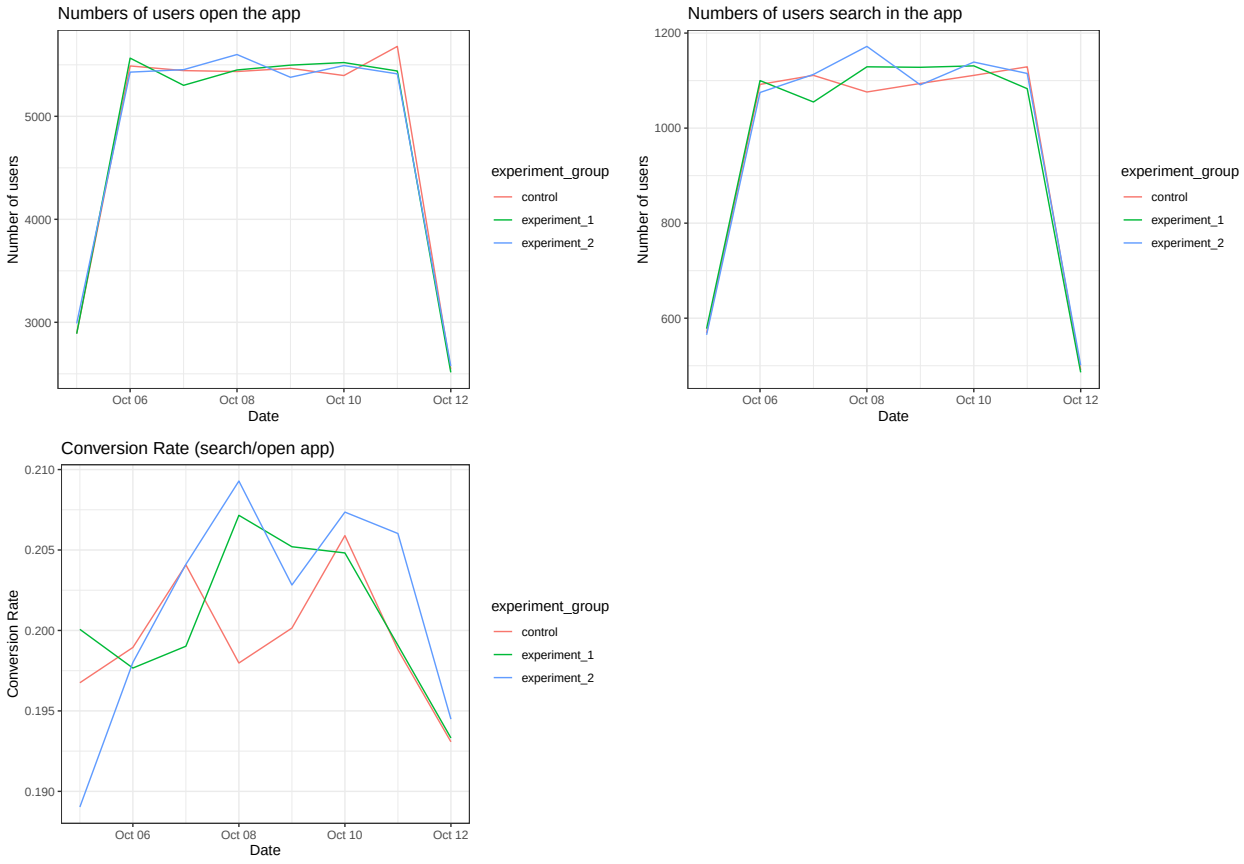
```
## `summarise()` has grouped output by 'experiment_group', 'event_day'. You can override using the `.gr
```

```
# number of open app by day
df1_open = df1 %>% filter(event_type == 'open')
ggplot(data=df1_open, aes(x=event_day, y=cnt, group=experiment_group)) +
```

```
  geom_line(aes(color=experiment_group)) +
  labs(title="Numbers of users open the app", x = "Date", y = "Number of users") +
  theme(plot.title=element_text(size=30)) +
  theme_bw()

# Number of search by day
df1_search = df1 %>% filter(event_type == 'search')
ggplot(data=df1_search, aes(x=event_day, y=cnt, group=experiment_group)) +
  labs(title="Numbers of users search in the app", x = "Date", y = "Number of users") +
  geom_line(aes(color=experiment_group)) + theme_bw()

# Conversion rate(search/open app) by day
df_conv_rate = df2 %>% group_by(experiment_group, event_day) %>%
  summarise(open = sum(open), of_users = sum(of_users),
            search = sum(search), begin_ride = sum(begin_ride)) %>%
  mutate(conv_rate = search/open)
```

```
## `summarise()` has grouped output by 'experiment_group'. You can override using the `.groups` argumen
```

```
ggplot(data=df_conv_rate, aes(x=event_day, y=conv_rate, group=experiment_group)) +
  geom_line(aes(color=experiment_group)) +
  labs(title="Conversion Rate (search/open app)", x = "Date", y = "Conversion Rate") +
  theme(plot.title=element_text(size=30)) +
  theme_bw()
```



7. Randomization check

The first step is to make sure that users are randomized in the treatment and control groups, on average there is no difference between the three groups on any characteristics other than treatment. That is, the three groups should be similar in all pre-treatment variables in terms of age and neighborhood. We performed t.test on the variables individually against the test variables. First, we performed t.test on the age and neighborhood variables between the control and experiment 1. Then, we performed t.test on the age and neighborhood variables between the control and experiment 2.

```
control = df2 %>% filter(experiment_group == 'control')
experiment1 = df2 %>% filter(experiment_group == 'experiment_1')
experiment2 = df2 %>% filter(experiment_group == 'experiment_2')
```

```
# Check randomization: control vs experiment 1
# User neighborhood
t.test(control$Bronx, experiment1$Bronx)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Bronx and experiment1$Bronx
## t = 0.49245, df = 76546, p-value = 0.6224
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001804100  0.003014853
## sample estimates:
##  mean of x  mean of y
## 0.03011263 0.02950725
```

```
t.test(control$Brooklyn, experiment1$Brooklyn)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Brooklyn and experiment1$Brooklyn
## t = -1.4902, df = 76537, p-value = 0.1362
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.010037919  0.001366965
## sample estimates:
## mean of x mean of y
## 0.2011680 0.2055035
```

```
t.test(control$Manhattan, experiment1$Manhattan)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Manhattan and experiment1$Manhattan
## t = 0.25143, df = 76546, p-value = 0.8015
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.005671196  0.007340327
## sample estimates:
## mean of x mean of y
## 0.6983001 0.6974656
```

```
t.test(control$Queens, experiment1$Queens)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Queens and experiment1$Queens
## t = 1.0003, df = 76539, p-value = 0.3172
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001503719  0.004638524
## sample estimates:
##  mean of x  mean of y
## 0.05021379 0.04864638
```

```
t.test(control$StatenIsland, experiment1$StatenIsland)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$StatenIsland and experiment1$StatenIsland
## t = 1.3274, df = 76483, p-value = 0.1844
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.000632927  0.003289193
## sample estimates:
##  mean of x  mean of y
## 0.02020544 0.01887731
```

```
# Age: f18t29, f30t39, f40t49, f50
t.test(control$f18t29, experiment1$f18t29)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f18t29 and experiment1$f18t29
## t = -0.30648, df = 76545, p-value = 0.7592
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.005845917  0.004264898
## sample estimates:
## mean of x mean of y
## 0.1493378 0.1501283
```

```
t.test(control$f30t39, experiment1$f30t39)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f30t39 and experiment1$f30t39
## t = -0.94185, df = 76534, p-value = 0.3463
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.006296395  0.002209164
## sample estimates:
##  mean of x  mean of y
## 0.09909792 0.10114154
```

```
t.test(control$f40t49, experiment1$f40t49)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f40t49 and experiment1$f40t49
## t = -2.4363, df = 76532, p-value = 0.01484
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.013730693 -0.001487482
## sample estimates:
## mean of x mean of y
## 0.2445771 0.2521862
```

```
t.test(control$f50, experiment1$f50)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f50 and experiment1$f50
## t = 2.8895, df = 76547, p-value = 0.003859
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.003359472 0.017526954
## sample estimates:
## mean of x mean of y
## 0.5069872 0.4965440
```

```
# Check randomization: control vs experiment 2
# User neighborhood
t.test(control$Bronx, experiment2$Bronx)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Bronx and experiment2$Bronx
## t = 0.18224, df = 76696, p-value = 0.8554
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002190083  0.002639111
```

```
## sample estimates:
##  mean of x  mean of y
## 0.03011263 0.02988812
```

```
t.test(control$Brooklyn, experiment2$Brooklyn)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Brooklyn and experiment2$Brooklyn
## t = -0.050579, df = 76697, p-value = 0.9597
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.005821396  0.005528504
## sample estimates:
## mean of x mean of y
## 0.2011680 0.2013145
```

```
t.test(control$Manhattan, experiment2$Manhattan)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Manhattan and experiment2$Manhattan
## t = -0.4888, df = 76697, p-value = 0.625
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.008110915  0.004872885
## sample estimates:
## mean of x mean of y
## 0.6983001 0.6999192
```

```
t.test(control$Queens, experiment2$Queens)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$Queens and experiment2$Queens
## t = 0.20449, df = 76696, p-value = 0.838
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002764431  0.003408471
## sample estimates:
##  mean of x  mean of y
## 0.05021379 0.04989177
```

```
t.test(control$StatenIsland, experiment2$StatenIsland)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  control$StatenIsland and experiment2$StatenIsland
## t = 1.2177, df = 76627, p-value = 0.2233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.000742971  0.003180826
## sample estimates:
##  mean of x  mean of y
## 0.02020544 0.01898652
```

```
# Age: f18t29, f30t39, f40t49, f50
t.test(control$f18t29, experiment2$f18t29)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f18t29 and experiment2$f18t29
## t = -0.41422, df = 76696, p-value = 0.6787
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.006120163  0.003984629
## sample estimates:
## mean of x mean of y
## 0.1493378 0.1504055
```

```
t.test(control$f30t39, experiment2$f30t39)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f30t39 and experiment2$f30t39
## t = 0.19008, df = 76697, p-value = 0.8493
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.003815620  0.004635156
## sample estimates:
##  mean of x  mean of y
## 0.09909792 0.09868816
```

```
t.test(control$f40t49, experiment2$f40t49)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f40t49 and experiment2$f40t49
## t = -1.041, df = 76695, p-value = 0.2979
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.009336303  0.002859006
## sample estimates:
## mean of x mean of y
## 0.2445771 0.2478158
```

```
t.test(control$f50, experiment2$f50)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$f50 and experiment2$f50
## t = 1.0792, df = 76697, p-value = 0.2805
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.003180209  0.010973505
## sample estimates:
## mean of x mean of y
## 0.5069872 0.5030905
```

Interpretation:

- Null hypothesis: There is no difference in the age/neighborhood between the control and experiment 1/2.
- Alternative hypothesis: There is an difference in the age/neighborhood between the control and experiment 1/2.

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

We could see that from the above t-tests, 89% of p-values are greater than 0.05. Thus, we failed to reject the null hypothesis: the age/neighborhood are not different between the control and experiment 1/2. Therefore, the users in the control and treatment groups are randomized.

8. Sample size check

We expected that experiments 1 and 2 will increase the conversion rate by 1%. Therefore, the next statistical test should check whether the available sample size is sufficient reliably to detect the difference between the treatment and control groups. To check the sufficiency of sample size, we used the power_t_test function and specify the arguments required.

```
mean(df2$search)
```

```
## [1] 0.2013526
```

```
sd(df2$search)
```

```
## [1] 0.4010126
```

```
table(df2$experiment_group)
```

```
##
##      control experiment_1 experiment_2
##        38356        38194        38343
```

```
power.t.test(n=38194, power=.8, sig.level=0.05, sd=0.4)
```

```
##
##      Two-sample t test power calculation
##
##               n = 38194
##           delta = 0.008115087
##              sd = 0.4
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Interpretation:

From the result of the test, the current sample size is only able to detect a difference of 0.008. Therefore, our experiment appears to be underpowered to detect the effect that management is looking for. Thus, it is recommended to re-run the experiment with a larger sample data set.

9. Experiment

Hypothesis 1: Redesign the text on the booking button would lead to more conversions (Control vs experiment 1)

To check whether Experiment 1 group has more conversions than the control group, we used t-test to testify the hypothesis.

```
t.test(control$search, experiment1$search, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$search and experiment1$search
## t = -0.4468, df = 76545, p-value = 0.655
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.006968298  0.004381106
## sample estimates:
## mean of x mean of y
## 0.2000469 0.2013405
```

Interpretation:

From the result, since the p-value is greater than 0.05, we failed to reject the null hypothesis: the conversion is not different between the control and experiment 1.

Hypothesis 2: Remove the "tip included" text would lead to more conversions (Control vs experiment 2)

To check whether Experiment 2 group has more conversions than the control group, we used t-test to testify the hypothesis.

```
t.test(control$search, experiment2$search, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  control$search and experiment2$search
## t = -0.90597, df = 76695, p-value = 0.365
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.008299864  0.003052460
## sample estimates:
## mean of x mean of y
## 0.2000469 0.2026706
```

Interpretation:

From the result, the p-value is also greater than 0.05, we failed to reject the null hypothesis: the conversion is not different between the control and experiment 2.

10. Conclusion

Based on the experimental results above, we can draw the following conclusions: Neither redesigning the text on the booking button nor removing the "tip included" text would increase the conversion rate between the "Open" and "Search" stages of the user journey. That is, there is no difference between the control and treatment groups. As a result, our team advises the Flyber not to make changes to these two features on their UI. Our team recommends that the company should conduct Usability Testing first, observe user interactions with the booking page, and then find out the root cause for the poor conversion rate of the booking page.

11. Limitation

Even though the A/B analysis did not show a promising improvement in conversions, this might be the funnel impact that will take a long term to get into effect, because the KPI requires long time periods and a very large sample to test for reliability. However, digging deeper would also create another problem when we try to speed up the experiment, developers may push visitors to the three experiment UI in a manner that over-resembles that natural flow on the website. As a result, the data will be skewed. Therefore, we should make sure the traffic to the experiment pages followed the same path during and after the A/B analysis.