AB Testing Result Analysis

Jinxin Ma 01/24/2018
jma33@dons.usfca.edu

About me

Education:

- MS in Data Science
- MA in Applied Economics
- BA in Economics

Experiences:

- Research Assistant at USF
- Data Science Intern at Silicon Valley Bank
- Risk Data Science Intern at Xoom – PayPal
- Research Analyst Co-op at BRA

About me

CONTENTS

Problem Statement

1. At company XYZ, France-based users have a much higher conversion rate than any other French-speaking country
2. The solution was to have one translation written by a local, aiming to improve the conversion rate
3. However, based on the test the localized translation was doing worse!

Is the test result really negative?

Data

1. Test data (453,321 users):
    - Conversion: whether a user converted (1) or not (0)
    - Test: whether the user is in test (1) or control (0)
    - Device: device used by the user
    - Ads Channel: marketing channel (FB, Google, etc)
    - User ID
    - etc.

2. User data (452,867 users): sex, age, country, user ID

Data (cont'd)

A quick glimpse of the data:

user_id	date	source	device	ads_channel	browser	sex	age	country	conversion	test
315281	12/3/15	Direct	Web		IE	M	32	Canada	**1**	**0**
497851	12/4/15	Ads	Web	Google	IE	M	21	Algeria	**1**	**1**
848402	12/4/15	Ads	Web	Facebook	Chrome	M	34	Canada	**0**	**0**
290051	12/3/15	Ads	Mobile	Facebook	Android_App	F	22	Algeria	**0**	**1**
548435	11/30/15	Ads	Web	Google	FireFox	M	19	Andorra	**0**	**1**

Data (cont'd)

Columns with missing values:

- Ads channel
- Sex
- Age
- Country

Use "NA" to fill missing values for ads channel, sex, and country

Use median to fill missing values for age

Variable	Missing value count
user_id	0
date	0
source	0
device	0
browser_language	0
ads_channel	**271444**
browser	0
conversion	0
test	0
sex	**454**
age	**454**
country	**454**

Problem Investigation

Conversion rate: treatment 4.3%, control 4.8%

Assumption:

$$H_0: Conversion\ rate_{treatment\ group} = Conversion\ rate_{control\ group}$$

$$H_1: Conversion\ rate_{treatment\ group} < Conversion\ rate_{control\ group}$$

Problem Investigation (cont'd)

Use a two-sample Z-test to check if conversion rate difference is statistically significant between two groups

$$\text{Z- score} = \frac{(\overline{p}_1 - \overline{p}_2) - 0}{\sqrt{\overline{p}(1-\overline{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Z-score: -7.43

P-value: 5.56 x 10^{-14}

$H_0: Conversion\ rate_{treatment\ group} = Conversion\ rate_{control\ group}$ REJECTED

$H_1: Conversion\ rate_{treatment\ group} < Conversion\ rate_{control\ group}$

Problem Investigation (cont'd)

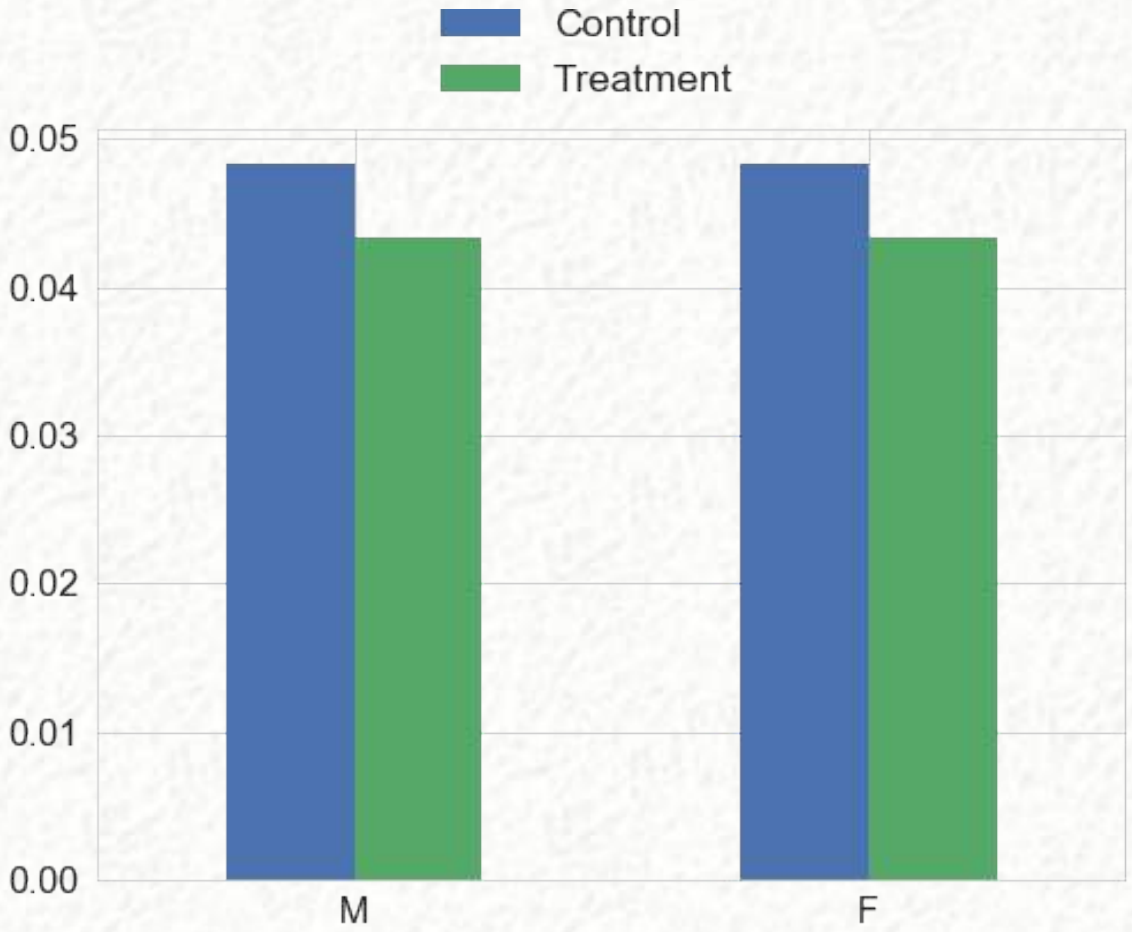
Is there any confounding variable?

A **confounding variable** is an outside influence that changes the effect of an independent variable on the dependent variable

Problem Investigation (cont'd)

Conversion rate by sex

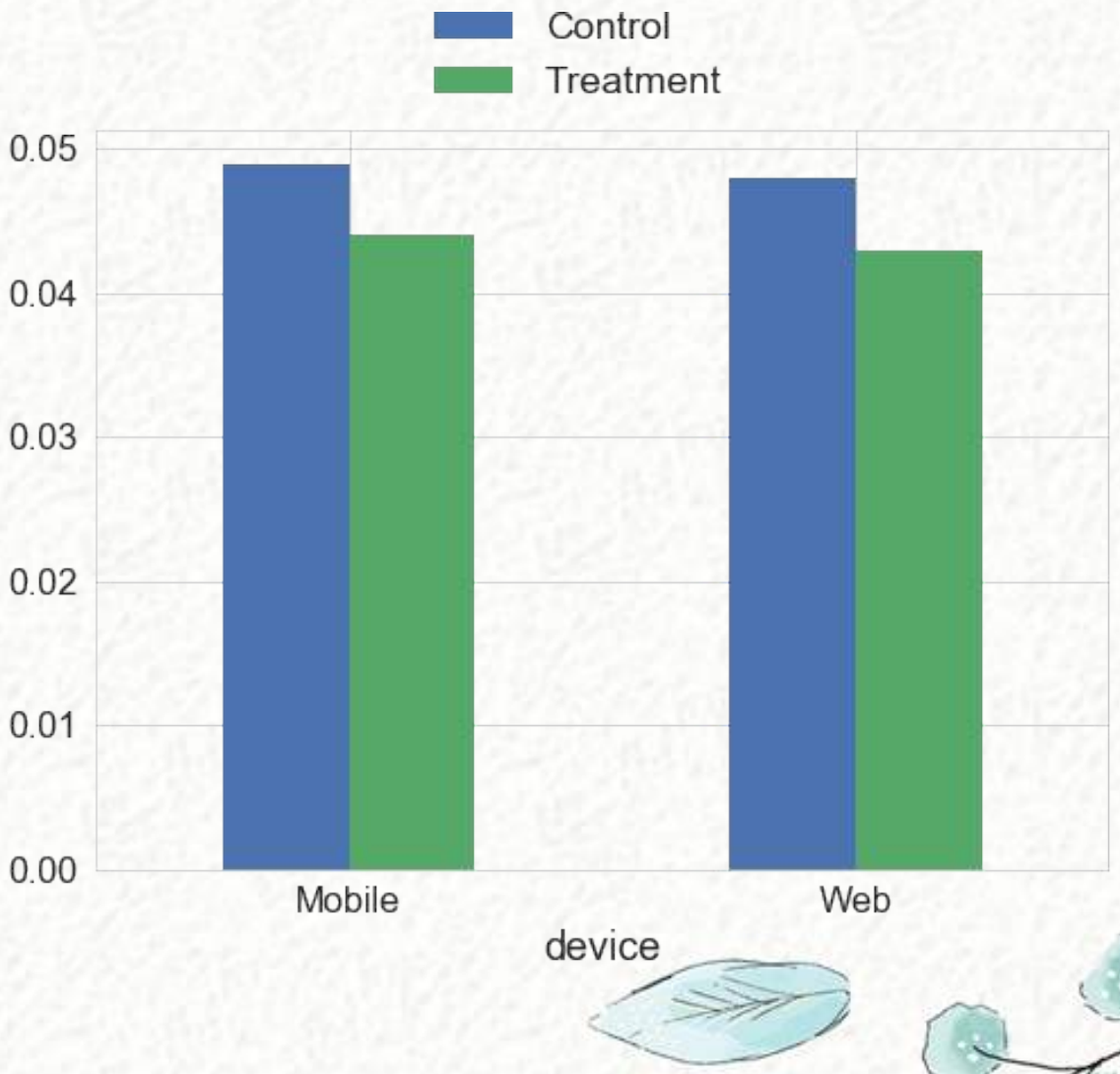
Sex doesn't seem to affect conversion rates



Problem Investigation (cont'd)

Conversion rate by device

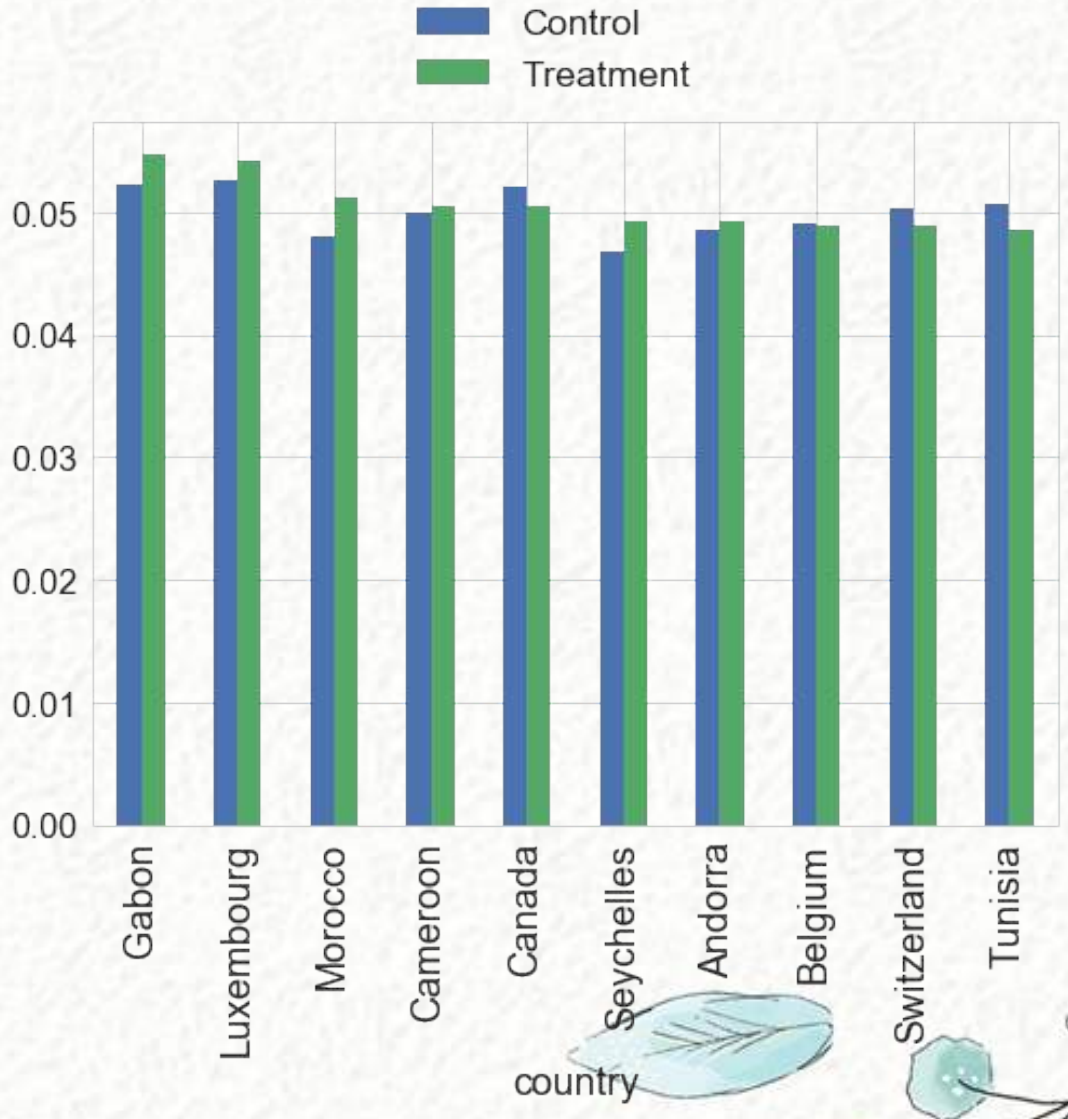
No effect from device



Problem Investigation (cont'd)

Conversion rate by country

Some countries have higher conversions in the treatment group!



Problem Investigation (cont'd)

Perform a two-sample Z-test for each country

$H_0: Conversion_{treatment} = Conversion_{control}$

$H_1: Conversion_{treatment} < Conversion_{control}$

Localized translation didn't change the conversion rate

Country	P-value
Republic of the Congo	0.124
Algeria	0.161
Canada	0.212
Senegal	0.236
Tunisia	0.280
Switzerland	0.287
Haiti	0.359
Belgium	0.481
Andorra	0.558
Mauritius	0.559

Solution

How do we control for the effect of **confounding variables** such as country so the same problem won't happen again?

Solution 1 - Stratification

Stratification is the process of dividing members of the population into homogeneous subgroups.

In our study, we can group users by country and/or other variables and perform statistical tests within each sub-population to check if conversion rates are statistically different between the treatment and the control groups.

Solution 1 – Pseudo code

Run a Z-test on the entire dataset, if p-value < 0.05:

- Put data points into **N** strata based on some variable
- Perform two-sample Z-tests (treatment vs. control) within each group and we will have **N** p-values
- If proportion of large p-values (>0.05) is less than some threshold (0.95 or 0.9), return false. Else, return true

This method works well when we have few strata and lots of data

Solution 1 – Result

All p-values are greater than 0.05

Country	P-value
Republic of the Congo	0.124
Algeria	0.161
Canada	0.212
Senegal	0.236
Tunisia	0.280
Switzerland	0.287
Haiti	0.359
Belgium	0.481
Andorra	0.558
Mauritius	0.559

Solution 2 - Regression

We can also use multivariate **logistic regression** to control for the effect of confounding variables.

Regression 1:

$$conversion = b0 + b1 * test$$

Regression 2:

$$conversion = b0 + b1 * test + b2 * country$$

If $b1$ in regression 2 changes by over 10%, **return false**

Solution 2 – Pseudo code

1. Run a logistic regression with “test”, save coefficient
2. For each potential confounding variable **Var**:
    - Run regression $conversion = b0 + b1 * test + b2 * \textbf{Var}$
    - If $b1$ changed by over 10% **return false**
3. Run a regression with all variables and if $b1$ changed by over 10% **return false**
4. If none of the above conditions are met, return true

This method works well when we have several confounding variables but not lots of data

Solution 2 – Result

1. For logistic regression with only “test”, the coefficient $b1$ is -0.1122 with a p-value of < 0.005
2. For regression $conversion = b0 + b1 * test + b2 * \text{country}$, $b1$ is 0.0031 with a p-value of 0.837. $b1$ changed by over 10%

Conclusion

1. We are able to identify the root cause to the negative test result, that is **the confounding effect of country**
2. To avoid the same problem from happening again, we now have two methodologies available: **stratification** and **logistic regression**

Next steps

1. Confirm the quality of localized translation
2. Check how users are assigned to the experiment and the control group. Algeria has 37,377 users in the treatment, but only 9,356 in the control
3. Use Bonferronni correction for p-values in the stratification method to avoid multiple testing problem
4. Use percentile or clustering to bin age into groups to check whether within each age group the conversion rate is different between the control and the treatment
5. Improve the algorithms and write tests to check for potential bugs

Thank you!