StepUp Analytics

Analysis of A/B Testing Results

Using R programming

Akash Kumar Gupta | Boda Mahender | Deepa Singh

Banaras Hindu University

Introduction

This dataset contains information collected by results of A/B testing of a company's website concerning number of conversions on their website in different countries. We have two datasets named "ab_data" and "countries". These two datasets have information on 7 variables/attributes on 2,94,478 observations. The variables used in the data are described below: -

1) user_id = unique user_id of different users.
2) Timestamp = timestamp when the user visited the website.
3) Group = which group the unit is from treatment / control group.
4) landing_page = page visited by the user new_page/ old_page.
5) converted = whether a user converted or not (purchased the product or not).
6) country = the country from where user interacted with website.

The Objective of this analysis is to find which page works better in sense of conversion and also to find which page works better country wise i.e. to suggest if we should replace old_page from new_page or not.

First, we need to install necessary packages

```
> install.packages("dplyr")
> library(dplyr)

> install.packages("ggplot2")
> library(ggplot2)
```

Then we need to set our working directory for the project

```
> setwd("D:/Current_Project")
```

Now we import our datasets

```
> ab_data = read.csv("ab_data.csv")
> head(ab_data)
```

```
  user_id                   timestamp     group landing_page converted
1  851104 2017-01-21 22:11:48.556739   control     old_page         0
2  804228 2017-01-12 08:01:45.159739   control     old_page         0
3  661590 2017-01-11 16:55:06.154213 treatment     new_page         0
4  853541 2017-01-08 18:28:03.143765 treatment     new_page         0
5  864975 2017-01-21 01:52:26.210827   control     old_page         1
6  936923 2017-01-10 15:20:49.083499   control     old_page         0
```

```
> cont = read.csv("countries.csv")
> head(cont)
```

```
  user_id country
1  834778      UK
2  928468      US
3  822059      UK
4  711597      UK
5  710616      UK
6  909908      UK
```

To start working on our datasets we need to merge these two datasets by using the common variable "user_id"

```
> data0 = merge(ab_data, cont, by.x = "user_id",by.y = "user_id",all=TRUE)

> head(data0)
  user_id                   timestamp     group landing_page converted country
1  630000 2017-01-19 06:26:06.548941 treatment     new_page         0      US
2  630001 2017-01-16 03:16:42.560309 treatment     new_page         1      US
3  630002 2017-01-19 19:20:56.438330   control     old_page         0      US
4  630003 2017-01-12 10:09:31.510471 treatment     new_page         0      US
5  630004 2017-01-18 20:23:58.824994 treatment     new_page         0      US
6  630005 2017-01-17 21:22:25.940766 treatment     new_page         1      US
```

The total number of rows in the data we have

```
> nrow(data0)
[1] 294478
```

Total number of unique users

```
> nrow(distinct(data0, user_id, .keep_all = TRUE))
[1] 290584
```

At first, we need to remove the data which decreases the accuracy of our analysis.
Here we have 2 data rows of some users which creates confusion whether this user received new_page or old_page.

```
> temp3 = filter(data0, (group == "treatment"& landing_page == "new_page") |
(group == "control"& landing_page == "old_page"))

> nrow(temp3)
[1] 290585
```

By using this method of removing irrelevant information only those users which has irregular combinations, we are removing 1 data row of such user who has 2 data rows.
Let's check if there is any user_id left with 2 data rows in temp3?

```
> B = temp3[duplicated(temp3$user_id),]$user_id
```

The duplicate user_id index may change from data to data so storing duplicate user_id values in variable B

```
> temp3[temp3$user_id == B,]

       user_id            timestamp       group    landing_page converted cou
ntry
131713  773192 2017-01-14 02:55:59.590927 treatment  new_page         0    US
131714  773192 2017-01-09 05:37:58.781806 treatment  new_page         0    US
```

Here we found that there is one user which has combination of group =" treatment" and landing_page= "new_page".
We must remove one row from this one, because we believe that according to sampling methods simple random sampling without replacement is better than simple random sampling with replacement.

```
> A = which(temp3$user_id == B,arr.ind = TRUE) #STORING THE INDEX VALUES TO A VARIABLE A
> A
[1] 131713 131714

> data1 = temp3[-A[1:(length(A)-1)], ]
#THE INDEX MAY CHANGE FROM ONE SYSTEM TO OTHER SO WE USE
```

After removing the data which needs to be removed, we have to check for "NA" values or if there is any missing data.

```
> anyNA(data1)
[1] FALSE
```

Now we can work with our data
The probability of conversion regardless of page is

```
> Cnvrt_Prob = mean(data1$converted == 1)

> cat("The probability of an individual converting regardless of the page they receive is:",Cnvrt_Prob)

The probability of an individual converting regardless of the page they receive is: 0.1195971
```

The probability of conversion of treatment group as converted column has binary number (0's and 1's) we can use mean function

```
> temp5 = data1[data1$group == 'treatment',]

> TR_Cnvrt_Prob = mean(temp5$converted)
> TR_Cnvrt_Prob
[1] 0.1188081
```

The probability of conversion of control group

```
> temp6 = data1[data1$group == 'control',]

> CON_Cnvrt_Prob = mean(temp6$converted)
> CON_Cnvrt_Prob
[1] 0.1203863
```

Observed difference

```
> Obs_Diff = TR_Cnvrt_Prob-CON_Cnvrt_Prob
> Obs_Diff
[1] -0.001578239
```

Here we see that probability of conversion of both pages are almost equal (Not much difference) i.e. 0.001578239
TR_Cnvrt_ProB - 0.1188081
CON_Cnvrt_Prob - 0.1203863
So, there is no sufficient evidence present to say that new_page leads to more conversion.

HYPOTHSIS TESTING

Here we take null hypothesis as

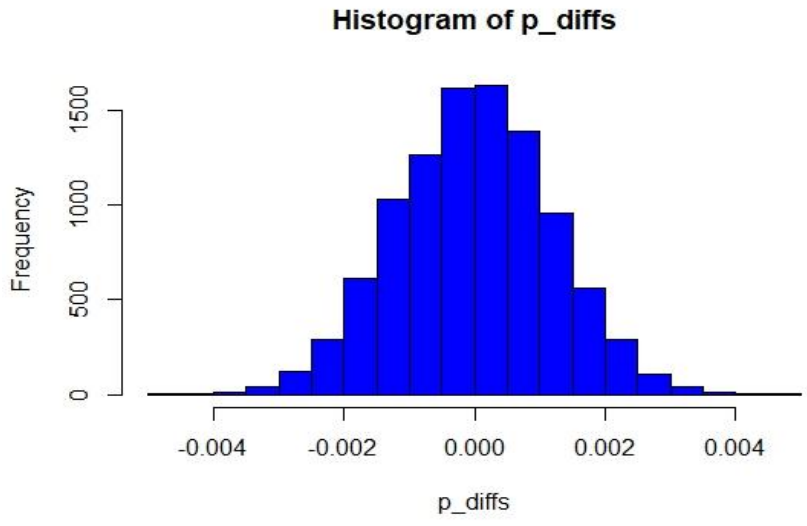
H_0 : TR_Cnvrt_Prob - CON_Cnvrt_Prob <= 0
H_1 : TR_Cnvrt_Prob - CON_Cnvrt_Prob0 > 0

We will assume that old_page is better unless the new_page proves to be definitely better at a type I error of **5%**
Also, we will assume that they are equal to Cnvrt_Prob

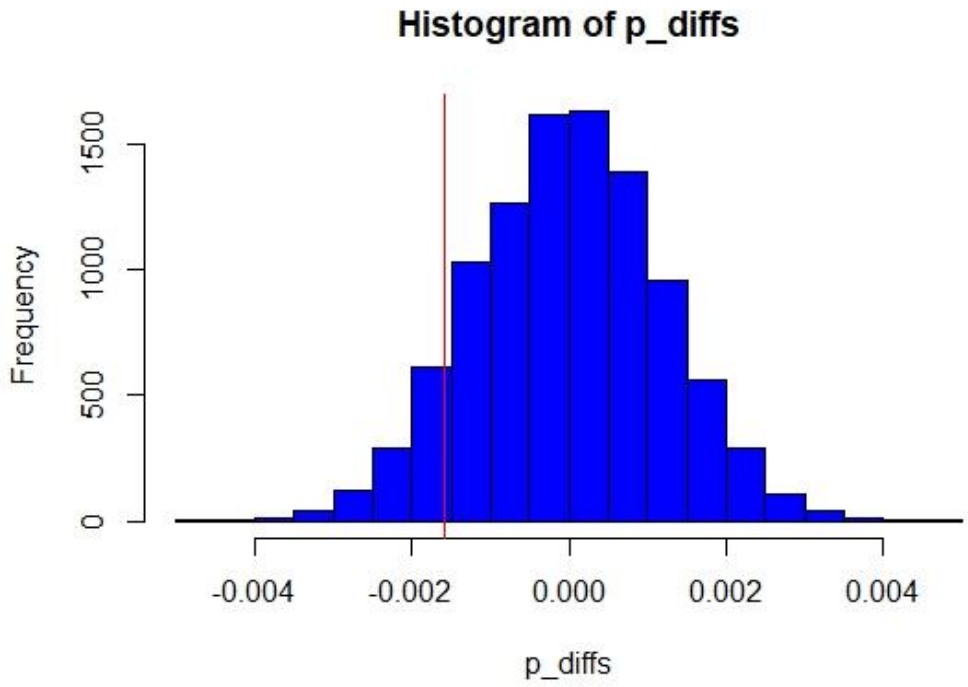
```
> p_new = Cnvrt_Prob
> p_old = Cnvrt_Prob

> n_New = nrow(filter(data1,landing_page == "new_page"))

> n_New
[1] 145310

> n_Old = nrow(filter(data1,landing_page == "old_page"))

> n_Old
[1] 145274
```

Stimulation of differences in conversion rates for null hypothesis

```
> p_diffs = 0> for (i in 1:10000)
+ {
+   new_page_cnrt = rbinom( n_New, size = 1, prob = c(p_new,1-p_new))
+   old_page_cnrt = rbinom( n_Old, size = 1, prob = c(p_old,1-p_old))
+   p_diffs = append(p_diffs, mean(new_page_cnrt) - mean(old_page_cnrt))
+ }

 > head(p_diffs)
 [1]  0.0000000000  0.0010666100  0.0004201588  0.0013079160 -0.0008878186
 [6] -0.0006265560

> hist( p_diffs,col = "blue")
```



```
> abline( v = Obs_Diff, col = "red")
```




```
> mean( p_diffs >= Obs_Diff )
[1] 0.9014099
```

Here our p value we get is 0.9062094 which is greater than 0.05 (our α)
So, we can't reject our null hypothesis.

We can also use functions to test possible rejection of our null hypothesis.

For this we will use prop.test()

```
> Num_Cnvrt_New = nrow(filter(data1,group == "treatment"& converted == 1))

> Num_Cnvrt_New
[1] 17264

> Num_Cnvrt_Old = nrow(filter(data1,group == "control"& converted == 1))

> Num_Cnvrt_Old
[1] 17489

> n_New
[1] 145310

> n_Old
[1] 145274

> prop.test(x = c( Num_Cnvrt_New, Num_Cnvrt_Old ), n = c( n_New, n_Old ),
+         p = NULL,alternative = "greater",conf.level = 0.95, correct = TRUE)
```

```
2-sample test for equality of proportions with continuity correction

data:  c(Num_Cnvrt_New, Num_Cnvrt_Old) out of c(n_New, n_Old)
X-squared = 1.7036, df = 1, p-value = 0.9041
alternative hypothesis: greater
95 percent confidence interval:
 -0.003565378  1.000000000
sample estimates:
   prop 1    prop 2
0.1188081 0.1203863
```

Here we ger p value 0.9041 which is almost equal to what value we got earlier.
So, we can't reject our null hypothesis.

Regression approach

Out of all variable "converted" is response variable (dependent) and other 6 variables are possible predi ctors (independent).

Since the variable **"converted"** has two values i.e. **0**'s and **1**'s, So we will use **logistic regression** for this.

For logistic regression the term logit is defined as

$$logit = log(p/1-p) = beta0 + beta1*X1 + error$$

First, we'll check for landing_page, if landing_page effects the conversion.

```
> model_1 = glm(converted ~ ab_page, family = binomial(link="logit"), data = data1)

> summary(model_1)

Call:
glm(formula = converted ~ ab_page, family = binomial(link = "logit"),
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5065  -0.5065  -0.5030  -0.5030   2.0641

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.988777   0.008062 -246.671   <2e-16 ***
ab_page     -0.014989   0.011434   -1.311     0.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212778  on 290583  degrees of freedom
Residual deviance: 212776  on 290582  degrees of freedom
AIC: 212780

Number of Fisher Scoring iterations: 4
```

The **p** value we get from this model is 0.190 which is still greater than 0.05 so we still can't reject the null Hypothesis H_0.

Now we'll check if variable "**country**" has any impact on conversion.
First, we need to check unique countries we have in our variable "country"

```
> unique(data1$country)
[1] US UK CA
Levels: CA UK US

> model_2 = glm(converted ~ country, family = binomial(link = "logit"),data = data1)

> summary(model_2)

Call:
glm(formula = converted ~ country, family = binomial(link = "logit"),
    data = data1)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5070  -0.5046  -0.5046  -0.5046   2.0785

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03753    0.02600 -78.365   <2e-16 ***
countryUK    0.05072    0.02839   1.786    0.074 .
countryUS    0.04080    0.02688   1.518    0.129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212778  on 290583  degrees of freedom
Residual deviance: 212775  on 290581  degrees of freedom
AIC: 212781

Number of Fisher Scoring iterations: 4
```

The **p** value we get from this model of different countries are 0.074 is 0.129 which is still greater than 0.05. So, we can say that country variable individually does not affect conversion.

Now, we'll check if any page works better in any particular country.

```
> data1$ab_page = ifelse(data1$landing_page == "new_page",1,0)
> View(data1)

> model_3 = glm(converted ~ country * ab_page, family = binomial(link = "logi
t"),data = data1)

> summary(model_3)

Call:
glm(formula = converted ~ country * ab_page, family = binomial(link = "logit"
),
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5083  -0.5071  -0.5057  -0.5022   2.0929

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.00401    0.03643 -55.008   <2e-16 ***
countryUK          0.01178    0.03984   0.296    0.767
countryUS          0.01753    0.03768   0.465    0.642
ab_page           -0.06745    0.05201  -1.297    0.195
countryUK:ab_page  0.07828    0.05680   1.378    0.168
countryUS:ab_page  0.04688    0.05378   0.872    0.383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212778  on 290583  degrees of freedom
Residual deviance: 212771  on 290578  degrees of freedom
AIC: 212783
Number of Fisher Scoring iterations: 4
```

The **p** value we get from this model of different interaction of landing_page and countries are still greate r than 0.05. So, we can say that interaction of landing_page and country does not affect conversion.

Conclusion

The Objective of this report is to suggest if the company should implement their new_page or not. Various statistical techniques were used to check if different variables affect the conversion like probability of conversion, hypothesis testing, two sample proportion test, logistics regression. We analyzed if new_page leads to more conversion and found that probability of conversion of both pages are equal.

We analyzed the effect of country variable on dependent variable “converted” and found that there is no significant effect of country on conversion individually. Countries do not influence significantly differences in the conversion rates.

And lastly, we checked if any page performs better in any particular country but found that there is no significant effect of this interaction of “landing_page” and “country” on conversion.

The convert rate may be related to some features of users like nationality, age, gender or specific cultural behavior. Adding additional information about users could reveal hidden value of the new version of the page for specific group of the users.

Acknowledgement

This A/B testing data analysis project is a golden opportunity for learning. We consider ourselves very lucky and honored to have so many wonderful people lead us through this attempt. Our grateful thanks to Mohammad Sajid and StepUp analytics team. They always helped us in every possible way whenever we faced a difficulty. We also thank our classmates and seniors for their support and solidarity. Thank you all.