Analyze_ab_test_results_notebook - Jupyter Notebook

Analyze A/B Test Results

This project will assure you have mastered the subjects covered in the statistics lessons. The hope is to have this project be as comprehensive of these topics as possible. Good luck!

Table of Contents

Introduction

A/B tests are very commonly performed by data analysts and data scientists. It is important that you get some practice working with the difficulties of these

For this project, you will be working to understand the results of an A/B test run by an e-commerce website. Your goal is to work through this notebook to help the company understand if they should implement the new page, keep the old page, or perhaps run the experiment longer to make their decision.

As you work through this notebook, follow along in the classroom and answer the corresponding quiz questions associated with each question. The labels for each classroom concept are provided for each question. This will assure you are on the right track as you work through the project, and you can feel more confident in your final submission meeting the criteria. As a final check, assure you meet all the criteria on the RUBRIC.

Part I - Probability

To get started, let's import our libraries.

Entrée [1]:

```
import pandas as pd

import numpy as np

import random

import matplotlib.pyplot as plt

%matplotlib inline

#We are setting the seed to assure you get the same answers on quizzes as we set up

random.seed(42)
```

`1.` Now, read in the `ab_data.csv` data. Store it in `df`. **Use your dataframe to answer the questions in Quiz 1 of the classroom.**

a. Read in the dataset and take a look at the top few rows here:

Entrée [2]:

```
df = pd.read_csv('ab_data.csv')

df.head()
```

Out[2]:

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

b. Use the below cell to find the number of rows in the dataset.

Entrée [3]:

```
df.shape[0]
```

Out[3]:

```
294478
```

c. The number of unique users in the dataset.

Entrée [4]:

```
df.user_id.nunique()
```

Out[4]:

```
290584
```

d. The proportion of users converted.

Entrée [5]:

```
df.converted.mean()*100
```

Out[5]:

```
11.96591935560551
```

e. The number of times the `new_page` and `treatment` don't line up.

Entrée [6]:

```
df[((df['group'] == 'treatment') == (df['landing_page'] == 'new_page')) == False].shape[0]
```

Out[6]:

```
3893
```

f. Do any of the rows have missing values?

Entrée [7]:

```
df.isnull().sum()
```

Out[7]:

```
user_id         0
timestamp       0
group           0
landing_page    0
converted       0
dtype: int64
```

`2.` For the rows where **treatment** is not aligned with **new_page** or **control** is not aligned with **old_page**, we cannot be sure if this row truly received the new or old page. Use **Quiz 2** in the classroom to provide how we should handle these rows.

a. Now use the answer to the quiz to create a new dataset that meets the specifications from the quiz. Store your new dataframe in **df2**.

Entrée [8]:

```
df2 = df.copy()
```

Entrée [9]:

```
df2.drop(df2.query('group=="control" & landing_page == "new_page"').index, inplace=True)
```

Entrée [10]:

```
df2.drop(df2.query('group=="treatment" & landing_page == "old_page"').index, inplace=True)
```

Entrée [11]:

```
df2.head(5)
```

Out[11]:

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

Entrée [12]:

```
# Double Check all of the correct rows were removed - this should be 0

df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) == False].shape[0]
```

Out[12]:

```
0
```

3. Use **df2** and the cells below to answer questions for **Quiz3** in the classroom.

a. How many unique **user_id**s are in **df2**?

Entrée [13]:

```
df2.user_id.nunique()
```

Out[13]:

```
290584
```

b. There is one **user_id** repeated in **df2**. What is it?

Entrée [14]:

```
df2.user_id.value_counts().head(1)
```

Out[14]:

```
773192    2
Name: user_id, dtype: int64
```

c. What is the row information for the repeat **user_id**?

Entrée [15]:

```
df2.query('user_id==773192')
```

Out[15]:

	user_id	timestamp	group	landing_page	converted
1899	773192	2017-01-09 05:37:58.781806	treatment	new_page	0
2893	773192	2017-01-14 02:55:59.590927	treatment	new_page	0

d. Remove **one** of the rows with a duplicate **user_id**, but keep your dataframe as **df2**.

Entrée [16]:

```
df2.drop(1899, inplace=True)

df2.query('user_id==773192')
```

Out[16]:

	user_id	timestamp	group	landing_page	converted
2893	773192	2017-01-14 02:55:59.590927	treatment	new_page	0

4. Use **df2** in the below cells to answer the quiz questions related to **Quiz 4** in the classroom.

a. What is the probability of an individual converting regardless of the page they receive?

Entrée [17]:

```
df2.converted.mean()*100
```

Out[17]:

```
11.959708724499627
```

b. Given that an individual was in the `control` group, what is the probability they converted?

Entrée [18]:

```
df.query('group=="control"').converted.mean()*100
```

Out[18]:

```
12.039917935897611
```

c. Given that an individual was in the `treatment` group, what is the probability they converted?

Entrée [19]:

```
df.query('group=="treatment"').converted.mean()*100
```

Out[19]:

```
11.891957956489856
```

d. What is the probability that an individual received the new page?

Entrée [20]:

```
df.query('landing_page=="new_page"').converted.mean()*100
```

Out[20]:

```
11.884079625642663
```

e. Consider your results from a. through d. above, and explain below whether you think there is sufficient evidence to say that the new treatment page leads to more conversions.

We see that the old page does better, but by a very small margin. Thus, we cannot say with certainty that a page leads to more conversions.

Part II - A/B Test

Notice that because of the time stamp associated with each event, you could technically run a hypothesis test continuously as each observation was observed.

However, then the hard question is do you stop as soon as one page is considered significantly better than another or does it need to happen consistently for a certain amount of time? How long do you run to render a decision that neither page is better than another?

These questions are the difficult parts associated with A/B tests in general.

`1.` For now, consider you need to make the decision just based on all the data provided. If you want to assume that the old page is better unless the new page proves to be definitely better at a Type I error rate of 5%, what should your null and alternative hypotheses be? You can state your hypothesis in terms of words or in terms of ***pold*****pold** and ***pnew*****pnew**, which are the converted rates for the old and new pages.

*H*0H0 : *pnew*pnew <= *pold*pold

*H*1H1 : *pnew*pnew > *pold*pold

`2.` Assume under the null hypothesis, *pnew*pnew and *pold*pold both have "true" success rates equal to the **converted** success rate regardless of page - that is *pnew*pnew and *pold*pold are equal. Furthermore, assume they are equal to the **converted** rate in **ab_data.csv** regardless of the page.

Use a sample size for each page equal to the ones in **ab_data.csv**.

Perform the sampling distribution for the difference in **converted** between the two pages over 10,000 iterations of calculating an estimate from the null.

Use the cells below to provide the necessary parts of this simulation. If this doesn't make complete sense right now, don't worry - you are going to work through the problems below to complete this problem. You can use **Quiz 5** in the classroom to make sure you are on the right track.

a. What is the **convert rate** for p_{new} under the null?

Entrée [21]:

```
p_new = df2.converted.mean()

p_new
```

Out[21]:

```
0.11959708724499628
```

b. What is the **convert rate** for p_{old} under the null?

Entrée [22]:

```
p_old = df2.converted.mean()

p_old
```

Out[22]:

```
0.11959708724499628
```

c. What is n_{new}?

Entrée [23]:

```
n_new = len(df2.query("group == 'treatment'"))

n_new
```

Out[23]:

```
145310
```

d. What is n_{old}?

Entrée [24]:

```
n_old = len(df2.query("group == 'control'"))

n_old
```

Out[24]:

```
145274
```

e. Simulate n_{new} transactions with a convert rate of p_{new} under the null. Store these n_{new} 1's and 0's in **new_page_converted**.

Entrée [25]:

```
new_page_converted  = np.random.choice([1,0], size = n_new, p = [p_new, (1- p_new)])

len(new_page_converted)
```

Out[25]:

```
145310
```

f. Simulate n_{old} transactions with a convert rate of p_{old} under the null. Store these n_{old} 1's and 0's in **old_page_converted**.

Entrée [26]:

```
old_page_converted = np.random.choice([1,0], size = n_old, p = [p_old, (1- p_old)])

len(old_page_converted)
```

Out[26]:

```
145274
```

g. Find *pnew*pnew - *pold*pold for your simulated values from part (e) and (f).

Entrée [27]:

```
new_page_converted = new_page_converted[:len(old_page_converted)]
```

Entrée [28]:

```
p_diffs = (new_page_converted/n_new) - (old_page_converted/n_old)

p_diffs
```

Out[28]:

```
array([0., 0., 0., ..., 0., 0., 0.])
```

h. Simulate 10,000 *pnew*pnew - *pold*pold values using this same process similarly to the one you calculated in parts **a. through g.** above. Store all 10,000 values in a numpy array called **p_diffs**.

Entrée [29]:

```
p_diffs = []


for _ in range(10000):

    new_page_converted = np.random.choice([1,0], size = n_new, p = [p_new, (1- p_new)]).mean()

    old_page_converted = np.random.choice([1,0], size = n_old, p = [p_old, (1- p_old)]).mean()

    diffs = new_page_converted - old_page_converted

    p_diffs.append(diffs)
```

Entrée [30]:

```
p_diffs = np.array(p_diffs)
```

i. Plot a histogram of the **p_diffs**. Does this plot look like what you expected? Use the matching problem in the classroom to assure you fully understand what was computed here.

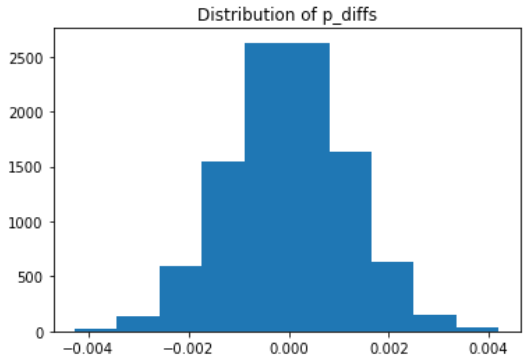
Entrée [31]:

```
plt.hist(p_diffs)

plt.title('Distribution of p_diffs')
```

Out[31]:

```
Text(0.5, 1.0, 'Distribution of p_diffs')
```

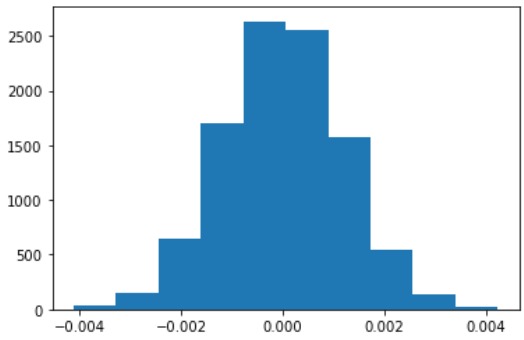


The P-diff is distributed normally and that's exactly what I expected. According to the Central Limit Theorem, when the sample size is large enough, the distribution of the sample mean is normally distributed.

Entrée [32]:

```
nullvals = np.random.normal(0, p_diffs.std(), 10000)

plt.hist(nullvals);
```



j. What proportion of the **p_diffs** are greater than the actual difference observed in **ab_data.csv**?

Entrée [33]:

```
actual_diff = df2.query("group == 'treatment'")['converted'].mean() - df2.query("group == 'control'")['converted'].mean()

actual_diff
```

Out[33]:

```
-0.0015782389853555567
```

Entrée [34]:

```
(nullvals > actual_diff).mean()
```

Out[34]:

```
0.9109
```

k. In words, explain what you just computed in part **j.** What is this value called in scientific studies? What does this value mean in terms of whether or not there is a difference between the new and old pages?

With a p-value of 0.90, we conclude that the difference in conversion rate between the new page and the old page does not appear to be significant. We fail to reject the null hypothesis that the new page is no better than the old page.

l. We could also use a built-in to achieve similar results. Though using the built-in might be easier to code, the above portions are a walkthrough of the ideas that are critical to correctly thinking about statistical significance. Fill in the below to calculate the number of conversions for each page, as well as the number of individuals who received each page. Let `n_old` and `n_new` refer the the number of rows associated with the old page and new pages, respectively.

Entrée [35]:

```
import statsmodels.api as sm


convert_old = sum(df2.query("group == 'control'")['converted'])

convert_new = sum(df2.query("group == 'treatment'")['converted'])

n_old = len(df2.query("group =='control'"))

n_new = len(df2.query("group =='treatment'"))
```

```
D:\Users\gaela\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:7: FutureWarning: pandas.Int64Index is
deprecated and will be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
  from pandas import (to_datetime, Int64Index, DatetimeIndex, Period,
D:\Users\gaela\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:7: FutureWarning: pandas.Float64Index is
deprecated and will be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
  from pandas import (to_datetime, Int64Index, DatetimeIndex, Period,
```

m. Now use `stats.proportions_ztest` to compute your test statistic and p-value. Here is a helpful link on using the built in.

Entrée [36]:

```
z_score, p_value = sm.stats.proportions_ztest([convert_old, convert_new], [n_old, n_new], alternative='smaller')

print(z_score, p_value)
```

```
1.3109241984234394 0.9050583127590245
```

n. What do the z-score and p-value you computed in the previous question mean for the conversion rates of the old and new pages? Do they agree with the findings in parts **j.** and **k.**?

The z-score and the p-value give the same message as the j and k parts, our p-value is very large, which implies that we fail to reject the null hypothesis. We then conclude that the new page is no better than the old page.

Part III - A regression approach

1. In this final part, you will see that the result you acheived in the previous A/B test can also be acheived by performing regression.

a. Since each row is either a conversion or no conversion, what type of regression should you be performing in this case?

Logistic regression cause of binary prediction.

b. The goal is to use **statsmodels** to fit the regression model you specified in part **a.** to see if there is a significant difference in conversion based on which page a customer receives. However, you first need to create a column for the intercept, and create a dummy variable column for which page each user received. Add an **intercept** column, as well as an **ab_page** column, which is 1 when an individual receives the **treatment** and 0 if **contracol**.

Entrée [37]:

```
df2.head(3)
```

Out[37]:

	user_id	timestamp	group	landing_page	converted
0	851104	2017-01-21 22:11:48.556739	control	old_page	0
1	804228	2017-01-12 08:01:45.159739	control	old_page	0
2	661590	2017-01-11 16:55:06.154213	treatment	new_page	0

Entrée [38]:

```
df2['intercept'] = 1


df2[['control', 'treatment']] = pd.get_dummies(df2['group'])

df2.head()
```

Out[38]:

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

c. Use **statsmodels** to import your regression model. Instantiate the model, and fit the model using the two columns you created in part **b.** to predict whether or not an individual converts.

Entrée [39]:

```
reg = sm.Logit(df2['converted'], df2[['intercept', 'treatment']])
```

d. Provide the summary of your model below, and use it as necessary to answer the following questions.

Entrée [40]:

```
results = reg.fit()

results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.366118
         Iterations 6
```

Out[40]:

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290582
Method:	MLE	**Df Model:**	1
Date:	Mon, 13 Jun 2022	**Pseudo R-squ.:**	8.077e-06
Time:	08:29:07	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.1899

Logit Regression Results

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-1.9888	0.008	-246.669	0.000	-2.005	-1.973
treatment	-0.0150	0.011	-1.311	0.190	-0.037	0.007

e. What is the p-value associated with **ab_page**? Why does it differ from the value you found in **Part II**?

Hint: What are the null and alternative hypotheses associated with your regression model, and how do they compare to the null and alternative hypotheses in the **Part II**?

The p-value associated with the treatment page here is different from Part II because the hypothesis test is different. Before we wanted to know if the conversion rate of the new page is higher than that of the old page but here we want to know if there is a difference in conversion rate between the new page and the old page. The P-value of 0.19 implies that we are not rejecting the null hypothesis, therefore, there is no difference in conversion rate between the new page and the old page. we have the same conclusions

f. Now, you are considering other things that might influence whether or not an individual converts. Discuss why it is a good idea to consider other factors to add into your regression model. Are there any disadvantages to adding additional terms into your regression model?

We can factor other factors into the regression model as they could also influence conversions. For example, user profiles (young, adult, old) which could create an aversion to change. This is a factor to consider when making the final decision. But let us know that by taking into account all these factors our logistic regression model will not necessarily be the most suitable.

g. Now along with testing if the conversion rate changes for different pages, also add an effect based on which country a user lives. You will need to read in the **countries.csv** dataset and merge together your datasets on the approporiate rows. Here are the docs for joining tables.

Does it appear that country had an impact on conversion? Don't forget to create dummy variables for these country columns - **Hint: You will need two columns for the three dummy variables.** Provide the statistical output as well as a written response to answer this question.

Entrée [41]:

```
countries = pd.read_csv('./countries.csv')

df3 = countries.set_index('user_id').join(df2.set_index('user_id'), how='inner')
```

Entrée [42]:

```
df3.head(3)
```

Out[42]:

	country	timestamp	group	landing_page	converted	intercept	control	treatment
user_id								
834778	UK	2017-01-14 23:08:43.304998	control	old_page	0	1	1	0
928468	US	2017-01-23 14:44:16.387854	treatment	new_page	0	1	0	1
822059	UK	2017-01-16 14:04:14.719771	treatment	new_page	1	1	0	1

Entrée [43]:

```
df3.country.value_counts()
```

Out[43]:

```
US    203619
UK     72466
CA     14499
Name: country, dtype: int64
```

Entrée [44]:

```
### Create the necessary dummy variables

df3[['CA', 'UK', 'US']] = pd.get_dummies(df3['country'])

df3.head(3)
```

Out[44]:

	country	timestamp	group	landing_page	converted	intercept	control	treatment	CA	UK	US
user_id											
834778	UK	2017-01-14 23:08:43.304998	control	old_page	0	1	1	0	0	1	0
928468	US	2017-01-23 14:44:16.387854	treatment	new_page	0	1	0	1	0	0	1
822059	UK	2017-01-16 14:04:14.719771	treatment	new_page	1	1	0	1	0	1	0

Entrée [45]:

```
df3['intercept'] = 1


reg2 = sm.Logit(df3['converted'], df3[['intercept', 'UK', 'US']])

results = reg2.fit()

results.summary()

Optimization terminated successfully.
         Current function value: 0.366116
         Iterations 6
```

Out[45]:

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290581
Method:	MLE	**Df Model:**	2
Date:	Mon, 13 Jun 2022	**Pseudo R-squ.:**	1.521e-05
Time:	08:29:12	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.1984

Logit Regression Results

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-2.0375	0.026	-78.364	0.000	-2.088	-1.987
UK	0.0507	0.028	1.786	0.074	-0.005	0.106
US	0.0408	0.027	1.518	0.129	-0.012	0.093

h. Though you have now looked at the individual factors of country and page on conversion, we would now like to look at an interaction between page and country to see if there significant effects on conversion. Create the necessary additional columns, and fit the new model.

Provide the summary results, and your conclusions based on the results.

Entrée [46]:

```
### Fit Your Linear Model And Obtain the Results


reg3 = sm.Logit(df3['converted'], df3[['intercept', 'treatment', 'UK', 'US']])

results2 = reg3.fit()

results2.summary()

Optimization terminated successfully.
         Current function value: 0.366113
         Iterations 6
```

Out[46]:

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290580
Method:	MLE	**Df Model:**	3
Date:	Mon, 13 Jun 2022	**Pseudo R-squ.:**	2.323e-05
Time:	08:29:16	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.1760

Logit Regression Results

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-2.0300	0.027	-76.249	0.000	-2.082	-1.978
treatment	-0.0149	0.011	-1.307	0.191	-0.037	0.007
UK	0.0506	0.028	1.784	0.074	-0.005	0.106
US	0.0408	0.027	1.516	0.130	-0.012	0.093

Even including both page and country in the template, there is still no significant effect on conversion.

Conclusions

The performance of the old page is slightly better, but we failed to reject the null hypothesis. The company should really consider either running the experiment longer or use their domain knowledge to identify some possible key characteristic of their best customers to help predict conversion rate.