Analyze_ab_test_results_notebook

March 17, 2018

0.1 Analyze A/B Test Results

You may either submit your notebook through the workspace here, or you may work from your local machine and submit through the next page. Either way assure that your code passes the project RUBRIC. **Please save regularly

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!

0.2 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.

```
In [52]: 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:

```
In [53]: df=pd.read_csv('ab_data.csv')
         df.head(3)
```

Out[53]:

	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

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

```
In [54]: len(df)
```

```
Out[54]: 294478
```

c. The number of unique users in the dataset.

```
In [55]: df['user_id'].nunique()
```

```
Out[55]: 290584
```

d. The proportion of users converted.

```
In [56]: df['converted'].mean()
```

```
Out[56]: 0.11965919355605512
```

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

```
In [57]: df_new=df.query("group!='treatment' and landing_page=='new_page'")
         len(df_new)
```

```
Out[57]: 1928
```

f. Do any of the rows have missing values?

```
In [58]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294478 entries, 0 to 294477
Data columns (total 5 columns):
user_id         294478 non-null int64
timestamp       294478 non-null object
group           294478 non-null object
landing_page    294478 non-null object
converted       294478 non-null int64
dtypes: int64(2), object(3)
memory usage: 11.2+ MB
```

Answer:- We can see that their no null values or missing value in any column

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**.

```
In [59]: df2=df.query("(group=='treatment' and landing_page=='new_page') or (group=='control' an
         df2.head()
```

Out[59]:

	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

```
In [60]: # Double Check all of the correct rows were removed - this should be 0
         df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) == False].sh
```

```
Out[60]: 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**?

```
In [61]: df2['user_id'].nunique()
```

```
Out[61]: 290584
```

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

```
In [62]: df2.set_index('user_id').index.get_duplicates()
```

```
Out[62]: [773192]
```

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

```
In [63]: df2.query("user_id=='773192'")
```

Out[63]:

	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**.

```
In [64]: df2=df2.drop_duplicates(['user_id'],keep='first')
         df2.query("user_id=='773192'")
```

Out[64]:

	user_id	timestamp	group	landing_page	converted
1899	773192	2017-01-09 05:37:58.781806	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?

```
In [65]: df2['converted'].mean()
```

```
Out[65]: 0.11959708724499628
```

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

```
In [66]: df2.query("group=='control'")['converted'].mean()
```

```
Out[66]: 0.1203863045004612
```

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

```
In [67]: df2.query("group=='treatment'")['converted'].mean()
```

```
Out[67]: 0.11880806551510564
```

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

```
In [68]: df_new=df2.query("landing_page=='new_page'")
         P=len(df_new)/len(df2)
         P
```

```
Out[68]: 0.5000619442226688
```

e. Use the results in the previous two portions of this question to suggest if you think there is evidence that one page leads to more conversions? Write your response below.

Answer:- yes the probability of group with control has more conversion rate which is 0.12038 and the probablity of group with treatment with conversion rate which is 0.11880.

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 p_{old} and p_{new}, which are the converted rates for the old and new pages.

Answer:-

Since Type 1 error rate =5%;

so alpha=0.05

Null hyphothesis=Old page is better than new page

Alternative hyphothesis=Old page is not better than new page

Pval<=0.05 ==>Reject Null hyphothesis ,Old page is not better than new page

Pval>0.05 ==>Accept Null hyphothesis,Old page is better than new page

2. Assume under the null hypothesis, p_{new} and p_{old} both have "true" success rates equal to the **converted** success rate regardless of page - that is p_{new} and p_{old} 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?

```
In [69]: p_new=df2['converted'].mean()
         p_new

Out[69]: 0.11959708724499628
```

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

```
In [70]: p_old=df2['converted'].mean()
         p_old

Out[70]: 0.11959708724499628
```

c. What is n_{new}?

```
In [71]: df_new=df2.query("landing_page=='new_page'")
         n_new=len(df_new)
         n_new

Out[71]: 145310
```

d. What is n_{old}?

```
In [72]: df_old=df2.query("landing_page=='old_page'")
         n_old=len(df_old)
         n_old

Out[72]: 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**.

```
In [73]: new_page_converted=np.random.choice([0,1],p=[(1-p_new),p_new],size=[n_new,1])
         type(new_page_converted),len(new_page_converted)
```

```
Out[73]: (numpy.ndarray, 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**.

```
In [74]: old_page_converted=np.random.choice([0,1],p=[1-p_old,p_old],size=[n_old,1])
         type(old_page_converted),len(old_page_converted)
```

```
Out[74]: (numpy.ndarray, 145274)
```

g. Find p_{new} - p_{old} for your simulated values from part (e) and (f).

```
In [75]: p_new - p_old
```

```
Out[75]: 0.0
```

h. Simulate 10,000 p_{new} - p_{old} values using this same process similarly to the one you calculated in parts **a. through g.** above. Store all 10,000 values in **p_diffs**.

```
In [76]: df_1=df2.query("converted=='1'")
         sample_len=len(df_1)
         sample_len
```

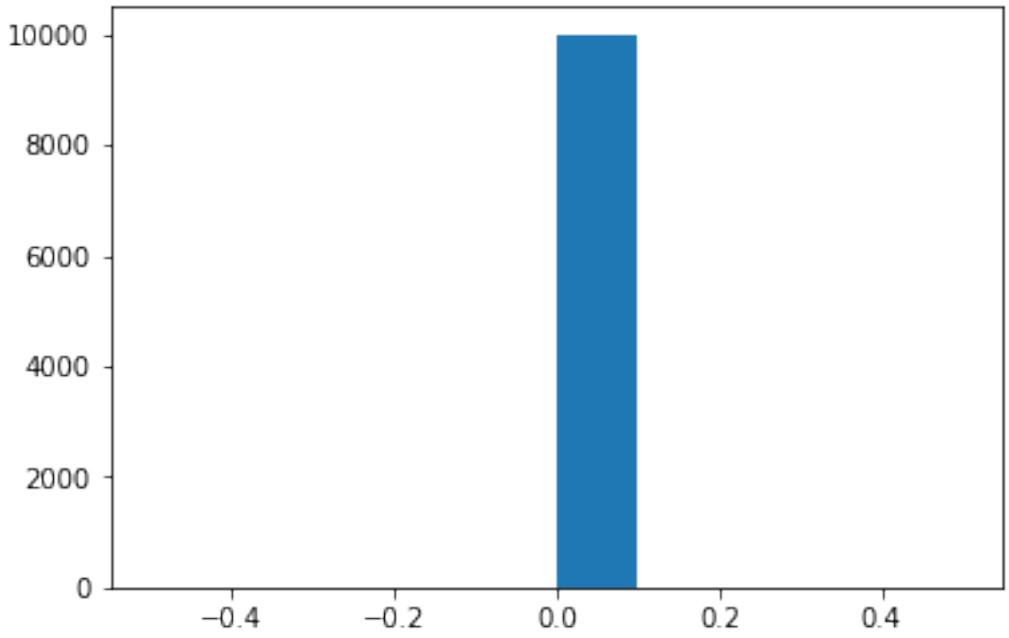
```
Out[76]: 34753
```

```
In [26]: p_diffs=[]
         for _ in range(10000):
             bootstrap_sample=df2.sample(sample_len,replace=True)
             p_new=bootstrap_sample['converted'].mean()
             p_old=bootstrap_sample['converted'].mean()
             p_diffs.append(p_new-p_old)
```

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.

```
In [29]: plt.hist(p_diffs)
```

```
Out[29]: (array([     0.,      0.,      0.,      0.,      0.,  10000.,      0.,
                      0.,      0.,      0.]),
          array([-0.5, -0.4, -0.3, -0.2, -0.1,  0. ,  0.1,  0.2,  0.3,  0.4,  0.5]),
          <a list of 10 Patch objects>)
```



Answer:-From the histogram we can understand p_diff is zero .Probability of conversion rate of new page and proability of conversion of old page is same

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

```
In [41]: cnt=0
         for s in p_diffs:
             if(s!=0):
                 print(s)
         print(cnt/len(p_diffs))
```

```
0.0
```

Answer:-It is same as previous one .Which means probability of conversion rate of old page and new page is same in calculated sample

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?

Answer:- we called it as 'p-value'.p-value determines significance of result.Since p-value is equal to zero .So it has strong evidence against the null hyphothesis.So old page is not better ,new page is better.We will reject null hyphothesis

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.

```
In [77]: import statsmodels.api as sm

         convert_old = len(df2.query("landing_page=='old_page' and converted=='1'"))
         convert_new = len(df2.query("landing_page=='new_page' and converted=='1'"))
         n_old = len(df2.query("landing_page=='old_page'"))
         n_new = len(df2.query("landing_page=='new_page'"))

In [78]: convert_old,convert_new,n_old,n_new

Out[78]: (17489, 17264, 145274, 145310)
```

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

```
In [79]: z_score, p_value = sm.stats.proportions_ztest([convert_old, convert_new], [n_old,n_new]
         z_score,p_value

Out[79]: (1.3109241984234394, 0.18988337448195103)
```

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.**?

Answer:- z-score is 1.3109 and p-value =0.18988 .We can see that p-value is high so we reject the null hyphotheis .So old page is not better than new page.

They don't agree finding in **j** and **k** .Since there value p-value is very less which is zero.So null hyphothesis is correct.Which means old page is better than new 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?

Answer:- Since it has two outcomes ,So we will use Logistic regression

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 colun 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 **control**.

```
In [80]: df2[['group_receive_1','ab_page']]=pd.get_dummies(df2['group'])

         df2.head()
```

Out[80]:

	user_id	timestamp	group	landing_page	converted	group_receive_1	ab_page
0	851104	2017-01-21 22:11:48.556739	control	old_page	0	1	0
1	804228	2017-01-12 08:01:45.159739	control	old_page	0	1	0
2	661590	2017-01-11 16:55:06.154213	treatment	new_page	0	0	1
3	853541	2017-01-08 18:28:03.143765	treatment	new_page	0	0	1
4	864975	2017-01-21 01:52:26.210827	control	old_page	1	1	0

```
In [81]: df2=df2.drop('group_receive_1',axis=1)
         df2.head()
```

Out[81]:

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

```
In [82]: df2['intercept']=1
         logit_mod=sm.Logit(df2['converted'],df2[['intercept','ab_page']])
         results=logit_mod.fit()

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

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

```
In [83]: results.summary()

Out[83]: <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:               290584
Model:                          Logit   Df Residuals:                   290582
Method:                           MLE   Df Model:                            1
Date:                Sat, 17 Mar 2018   Pseudo R-squ.:               8.077e-06
Time:                        01:17:34   Log-Likelihood:            -1.0639e+05
converged:                       True   LL-Null:                   -1.0639e+05
                                        LLR p-value:                    0.1899
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -1.9888      0.008   -246.669      0.000      -2.005      -1.973
ab_page       -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 the **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**?

Answer:- P-value = 0.190
Null hyphothesis:-Old page is better than new page
Alternative hyphothesis:Old page is not better than new page.
Since p-value is greater than 0.05.So we reject the null hyphothesis.
It means old page is not better than new page.
In part 2 ,P-value=0.119 ,Current P-value=0.190
Which is less than current value obtained through regression.
So logistic regression has more statistic significance.
In part 2 we have used A/B testing module .Here we have used logistic regression model which is best for the output which has two outcomes.

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?

Answer:- Adding new varibles in regression model means you are controlling constant for it ,if they are independent variable then the coefficient should not change.
Disavantage of adding additional term in regression model:- This usually comes down to the data being used.Examples of these are incomplete data being used and falsely concluding thant coorelation is a causation

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 varaibles.** Provide the statistical output as well as a written response to answer this question.

```
In [84]: df_c=pd.read_csv('countries.csv')
         df_c.head()
```

```
Out[84]:    user_id country
         0   834778      UK
         1   928468      US
         2   822059      UK
         3   711597      UK
         4   710616      UK
```

```
In [85]: df2=df2.join(df_c.set_index('user_id'),on='user_id')
         df2.head()
```

```
Out[85]:    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

            ab_page  intercept country
         0        0          1      US
         1        0          1      US
         2        1          1      US
         3        1          1      US
         4        0          1      US
```

```
In [86]: df2['country'].value_counts()
```

```
Out[86]: US    203619
         UK     72466
         CA     14499
         Name: country, dtype: int64
```

```
In [87]: df2[['cntry_US','cntry_UK','cntry_CA']]=pd.get_dummies(df2['country'])
         df2.head()
```

```
Out[87]:    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

            ab_page  intercept country  cntry_US  cntry_UK  cntry_CA
```

```
0        0          1      US          0          0          1
1        0          1      US          0          0          1
2        1          1      US          0          0          1
3        1          1      US          0          0          1
4        0          1      US          0          0          1
```

```
In [88]: df2=df2.drop('cntry_CA',axis=1)
```

Answer:-We can see it has three countries mainly 'US' ,'UK' and 'CA'.So we have created three dummy variable for three countries .We have dropped one column country 'CA' to make it full rank matrix

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.

```
In [89]: lm=sm.Logit(df2['converted'],df2[['intercept','cntry_US','cntry_UK']])
         results=lm.fit()
         results.summary()
```

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

```
Out[89]: <class 'statsmodels.iolib.summary.Summary'>
         """
                                  Logit Regression Results
         ==============================================================================
         Dep. Variable:              converted   No. Observations:               290584
         Model:                          Logit   Df Residuals:                   290581
         Method:                           MLE   Df Model:                            2
         Date:                Sat, 17 Mar 2018   Pseudo R-squ.:               1.521e-05
         Time:                        01:18:56   Log-Likelihood:            -1.0639e+05
         converged:                       True   LL-Null:                   -1.0639e+05
                                                 LLR p-value:                    0.1984
         ==============================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
         ------------------------------------------------------------------------------
         intercept     -1.9967      0.007   -292.314      0.000      -2.010      -1.983
         cntry_US      -0.0408      0.027     -1.518      0.129      -0.093       0.012
         cntry_UK       0.0099      0.013      0.746      0.456      -0.016       0.036
         ==============================================================================
         """
```

Answer:- US:- p-value=0.027 which is less than 0.05 .Which means null hyphothesis is accepted.So old page is better than new page. Uk:- p-value=0.013.which is less than 0.05 .Which means null hyphothesis is accepted.So old page is better than new page.

Conclusions

Congratulations on completing the project!

0.2.1 Gather Submission Materials

Once you are satisfied with the status of your Notebook, you should save it in a format that will make it easy for others to read. You can use the **File -> Download as -> HTML (.html)** menu to save your notebook as an .html file. If you are working locally and get an error about "No module name", then open a terminal and try installing the missing module using `pip install <module_name>` (don't include the "<" or ">" or any words following a period in the module name).

You will submit both your original Notebook and an HTML or PDF copy of the Notebook for review. There is no need for you to include any data files with your submission. If you made reference to other websites, books, and other resources to help you in solving tasks in the project, make sure that you document them. It is recommended that you either add a "Resources" section in a Markdown cell at the end of the Notebook report, or you can include a `readme.txt` file documenting your sources.

0.2.2 Submit the Project

When you're ready, click on the "Submit Project" button to go to the project submission page. You can submit your files as a .zip archive or you can link to a GitHub repository containing your project files. If you go with GitHub, note that your submission will be a snapshot of the linked repository at time of submission. It is recommended that you keep each project in a separate repository to avoid any potential confusion: if a reviewer gets multiple folders representing multiple projects, there might be confusion regarding what project is to be evaluated.

It can take us up to a week to grade the project, but in most cases it is much faster. You will get an email once your submission has been reviewed. If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com. In the meantime, you should feel free to continue on with your learning journey by continuing on to the next module in the program.