Analyze_ab_test_results_notebook

December 10, 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 [2]: 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 [3]: df = pd.read_csv('ab_data.csv')
        df.head()
```

Out[3]:

	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.

```
In [4]: len(df)
```

```
Out[4]: 294478
```

c. The number of unique users in the dataset.

```
In [5]: df.user_id.nunique()
```

```
Out[5]: 290584
```

d. The proportion of users converted.

```
In [6]: users = df.groupby('user_id').sum()
        number_unique_users = len(users)
        number_converted_users = users.converted.sum()
        number_converted_users / number_unique_users
```

```
Out[6]: 0.12126269856564711
```

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

```
In [7]: treatment_and_not_new_page = (df.group == 'treatment') & (df.landing_page != 'new_page')
        not_treatment_and_new_page = (df.group != 'treatment') & (df.landing_page == 'new_page')
        not_lined_up = treatment_and_not_new_page | not_treatment_and_new_page
        len(df[not_lined_up])
```

```
Out[7]: 3893
```

f. Do any of the rows have missing values?

```
In [8]: 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: No missing values.

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 [9]: treatment_and_new_page = (df.group == 'treatment') & (df.landing_page == 'new_page')
        control_and_old_page = (df.group == 'control') & (df.landing_page == 'old_page')
        clean_rows = control_and_old_page | treatment_and_new_page
        df2 = df[clean_rows]
```

```
In [10]: # 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[10]: 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 [11]: df2.user_id.nunique()
```

```
Out[11]: 290584
```

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

```
In [12]: userid = df2.groupby('user_id')['timestamp'].count().sort_values(ascending=False).head(
         userid
```

```
Out[12]: 773192
```

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

```
In [16]: df2[df2.user_id == userid]
```

```
Out[16]:       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 [13]: drop_row = df2[df2.user_id == userid].index[0]
         df2 = df2.drop(drop_row)
```

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 [14]: df2.converted.sum() / len(df2)

Out[14]: 0.11959708724499628
```

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

```
In [15]: control_group = df2.group == 'control'
         control_group_and_converted = control_group & (df2.converted == 1)
         len(df2[control_group_and_converted]) / len(df2[control_group])

Out[15]: 0.1203863045004612
```

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

```
In [16]: treatment_group = df2.group == 'treatment'
         treatment_group_and_converted = treatment_group & (df2.converted == 1)
         len(df2[treatment_group_and_converted]) / len(df2[treatment_group])

Out[16]: 0.11880806551510564
```

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

```
In [17]: len(df2[treatment_group]) / len(df2)

Out[17]: 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.

The control group has conversion rate of 12%

The treatment group has conversion rate of 11.9% The rate for control and tratment is very close therefore No any effect on conversion.

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.

Put your answer here. H1: Pnew-Pold>0

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 [18]: p_new_null = df2.converted.mean()
         p_new_null

Out[18]: 0.11959708724499628
```

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

```
In [19]: p_old_null = p_new_null
         p_old_null

Out[19]: 0.11959708724499628
```

c. What is n_{new}?

```
In [20]: n_new = (df2.landing_page == 'new_page').sum()
         n_new

Out[20]: 145310
```

d. What is n_{old}?

```
In [21]: n_old = (df2.landing_page == 'old_page').sum()
         n_old

Out[21]: 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 [56]: new_page_converted = np.random.choice([0,1],size=n_new,p=[(1-p_new_null),p_new_null])
```

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

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

```
In [58]: p_diff = new_page_converted.mean() - old_page_converted.mean()
         p_diff
```

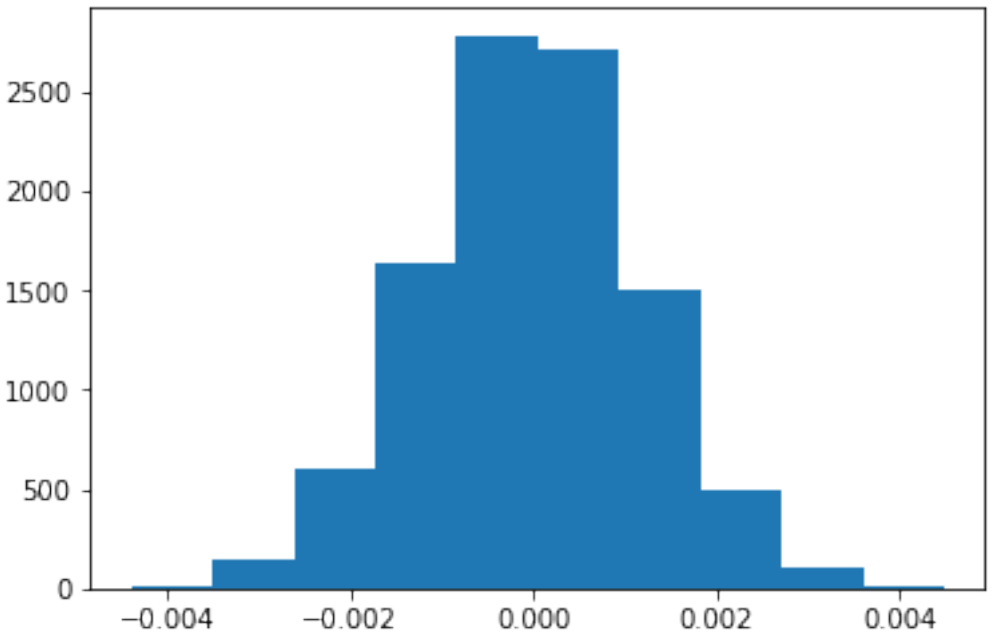
```
Out[58]: 0.00047945469922414108
```

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 [59]: p_diffs_values = []
         for _ in range(10000):
             new_page_converted = np.random.choice([0,1],size=n_new,p=[(1-p_new_null),p_new_null
             old_page_converted = np.random.choice([0,1],size=n_old,p=[(1-p_old_null),p_old_null
             p_diff = new_page_converted.mean() - old_page_converted.mean()
             p_diffs_values.append(p_diff)
         p_diffs = np.array(p_diffs_values)
```

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 [60]: plt.hist(p_diffs);
```



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

```
In [61]: p_old = df2[df2.landing_page == 'old_page'].converted.mean()
         p_new = df2[df2.landing_page == 'new_page'].converted.mean()
         p_diff_observed = p_new - p_old
         p_diff_observed

Out[61]: -0.0015782389853555567

In [62]: (p_diffs > p_diff_observed).mean()

Out[62]: 0.90480000000000005
```

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?

Put your answer here.
The p-value calculated is 0.904 is greater than the alpha 0.05.

```
In [63]: import statsmodels.api as sm

         convert_old = df2.query('group == "control" & converted == 1')['converted'].count()
         convert_new = df2.query('group == "treatment" & converted == 1')['converted'].count()
```

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 [64]: z_score, p_value = sm.stats.proportions_ztest([convert_old, convert_new], [n_old, n_new
         (z_score, p_value)

Out[64]: (1.3109241984234394, 0.90505831275902449)

In [65]: from scipy.stats import norm

         norm.cdf(z_score) # z-score

Out[65]: 0.90505831275902449

In [66]: norm.ppf(1-(0.05/2)) # critical value at 95% confidence

Out[66]: 1.959963984540054
```

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.**?

Put your answer here.
Assuming that we want 95% confidence in our conclusion, the z-score of 1.310 is not more than the critical value of 1.96 , reject H0 that the new page has a conversion rate 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?

Put your answer here. 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 [67]: df2[['ab_page', 'old_page']] = pd.get_dummies(df2['landing_page'])
         df2['intercept'] = 1
         df2.head()
```

Out[67]:

	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	old_page	intercept
0	0	1	1
1	0	1	1
2	1	0	1
3	1	0	1
4	0	1	1

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 [68]: from scipy import stats
         stats.chisqprob = lambda chisq, df: stats.chi2.sf(chisq, df)

In [69]: 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 [70]: results.summary()
```

```
Out[70]: <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:                Sun, 09 Dec 2018   Pseudo R-squ.:               8.077e-06
Time:                        11:07:45   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**?

Put your answer here. p-value with ab_page is 0.190.

This is greater than 0.05, we reject H0 and this is a difference in conversion rate based on which page the cusotmer receives.

In Part II, H0 of the old_page was greater than or equal. this is different in p-values in Part III we are see equal and not equal but in Part II, we were see less than or equal or greater than.

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?

Put your answer here.

It's better to think other factors to add to the model because to help us improve accuracy.

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 [71]: countries_df = pd.read_csv('countries.csv')
         df_new = countries_df.set_index('user_id').join(df2.set_index('user_id'), how='inner')

In [72]: # Create the necessary dummy variables
         df_new.country.unique()
```

```
Out[72]: array(['UK', 'US', 'CA'], dtype=object)
```

```
In [73]: df_new['intercept'] = 1
         df_new[['CA','UK','US']]=pd.get_dummies(df_new.country)
         df_new[['new_page','old_page']]=pd.get_dummies(df_new.landing_page)
         df_new.head()
```

Out[73]:

user_id	country	timestamp	group	landing_page	converted	ab_page	old_page	intercept	CA	UK	US	new_page
834778	UK	2017-01-14 23:08:43.304998	control	old_page	0	0	1	1	0	1	0	0
928468	US	2017-01-23 14:44:16.387854	treatment	new_page	0	1	0	1	0	0	1	1
822059	UK	2017-01-16 14:04:14.719771	treatment	new_page	1	1	0	1	0	1	0	1
711597	UK	2017-01-22 03:14:24.763511	control	old_page	0	0	1	1	0	1	0	0
710616	UK	2017-01-16 13:14:44.000513	treatment	new_page	0	1	0	1	0	1	0	1

```
In [74]: df_new['intercept'] = 1
         log_mod = sm.Logit(df_new.converted,df_new[['intercept','CA','UK','new_page']])
         results = log_mod.fit()
         results.summary()
```

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

```
Out[74]: <class 'statsmodels.iolib.summary.Summary'>
         """
                                  Logit Regression Results
         ==============================================================================
         Dep. Variable:              converted   No. Observations:               290584
         Model:                          Logit   Df Residuals:                   290580
         Method:                           MLE   Df Model:                            3
         Date:                Sun, 09 Dec 2018   Pseudo R-squ.:               2.323e-05
         Time:                        11:08:58   Log-Likelihood:            -1.0639e+05
         converged:                       True   LL-Null:                   -1.0639e+05
                                                 LLR p-value:                    0.1760
         ==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
         ------------------------------------------------------------------------------
         intercept     -1.9893      0.009   -223.763      0.000      -2.007      -1.972
         CA            -0.0408      0.027     -1.516      0.130      -0.093       0.012
```

```
UK             0.0099      0.013      0.743      0.457      -0.016       0.036
new_page      -0.0149      0.011     -1.307      0.191      -0.037       0.007
==============================================================================
"""
```

p-value with country dummy variables is greater than our alpha 0.05. The country is not important for this model.

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 [75]: df_new['CA_new_page'] = df_new.CA * df_new.new_page
         df_new['UK_new_page'] = df_new.UK * df_new.new_page
         df_new.head()
```

Out[75]:

user_id	country	timestamp	group	landing_page	converted	ab_page	old_page	intercept	CA	UK	US	new_page	CA_new_page	UK_new_page
834778	UK	2017-01-14 23:08:43.304998	control	old_page	0	0	1	1	0	1	0	0	0	0
928468	US	2017-01-23 14:44:16.387854	treatment	new_page	0	1	0	1	0	0	1	1	0	0
822059	UK	2017-01-16 14:04:14.719771	treatment	new_page	1	1	0	1	0	1	0	1	0	1
711597	UK	2017-01-22 03:14:24.763511	control	old_page	0	0	1	1	0	1	0	0	0	0
710616	UK	2017-01-16 13:14:44.000513	treatment	new_page	0	1	0	1	0	1	0	1	0	1

```
In [76]: log_mod = sm.Logit(df_new.converted,df_new[['intercept','CA','UK','new_page','CA_new_pa
         results = log_mod.fit()
         results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.366109
         Iterations 6
```

```
Out[76]: <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:               290584
Model:                          Logit   Df Residuals:                   290578
Method:                           MLE   Df Model:                            5
Date:                Sun, 09 Dec 2018   Pseudo R-squ.:               3.482e-05
Time:                        11:09:33   Log-Likelihood:            -1.0639e+05
converged:                       True   LL-Null:                   -1.0639e+05
                                        LLR p-value:                    0.1920
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      -1.9865      0.010   -206.344      0.000      -2.005      -1.968
CA             -0.0175      0.038     -0.465      0.642      -0.091       0.056
UK             -0.0057      0.019     -0.306      0.760      -0.043       0.031
new_page       -0.0206      0.014     -1.505      0.132      -0.047       0.006
CA_new_page    -0.0469      0.054     -0.872      0.383      -0.152       0.059
UK_new_page     0.0314      0.027      1.181      0.238      -0.021       0.084
==============================================================================
"""
```

p-value do not improve for any of the dummy variables. They are all still less than our α level of 0.05

Finishing Up

Congratulations! You have reached the end of the A/B Test Results project! This is the final project in Term 1. You should be very proud of all you have accomplished!

Tip: Once you are satisfied with your work here, check over your report to make sure that it is satisfies all the areas of the rubric (found on the project submission page at the end of the lesson). You should also probably remove all of the "Tips" like this one so that the presentation is as polished as possible.

0.3 Directions to Submit

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```
In [ ]: from subprocess import call
        call(['python', '-m', 'nbconvert', 'Analyze_ab_test_results_notebook.ipynb'])
```