Analyze_ab_test_results_notebook

December 8, 2020

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

```
In [4]: df.shape
```

```
Out[4]: (294478, 5)
```

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]: (df.converted.sum())/(df.shape[0])
```

```
Out[6]: 0.11965919355605512
```

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

```
In [6]: df[((df['group'] == 'treatment') != (df['landing_page'] == 'new_page')) == True].shape[0
```

```
Out[6]: 3893
```

f. Do any of the rows have missing values?

```
In [10]: 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
```

2. For the rows where **treatment** does not match with **new_page** or **control** does not match with **old_page**, we cannot be sure if this row truly received the new or old page. Use **Quiz 2** in the classroom to figure out 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 [7]: #Separating treatment and new page, control and old page, in two different dataframes th
        df2t = df.query('group == "treatment" and landing_page == "new_page"')
        df2c = df.query('group == "control" and landing_page == "old_page"')
        df2 = df2t.merge(df2c, how='outer')

In [8]: # Double Check all of the correct rows were removed - this should be 0
        df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) == False].sha

Out[8]: 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 [9]: df2.user_id.nunique()

Out[9]: 290584
```

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

```
In [10]: ids = df2.user_id
         df2[ids.isin(ids[ids.duplicated()])].user_id

Out[10]: 938     773192
         1404    773192
         Name: user_id, dtype: int64
```

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

```
In [11]: df2[ids.isin(ids[ids.duplicated()])]

Out[11]:       user_id                   timestamp      group landing_page  converted
         938    773192  2017-01-09 05:37:58.781806  treatment     new_page          0
         1404   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 [12]: df2 = df2[~df2.user_id.duplicated(keep='first')]
```

4. Use **df2** in the cells below 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 [13]: df2.converted.mean()
```

```
Out[13]: 0.11959708724499628
```

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

```
In [14]: df2c = df2.query('group == "control"')
         df2c.converted.mean()

Out[14]: 0.1203863045004612
```

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

```
In [15]: df2t = df2.query('group == "treatment"')
         df2t.converted.mean()

Out[15]: 0.11880806551510564
```

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

```
In [16]: len(df2t.index)/len(df2.index)

Out[16]: 0.5000619442226688
```

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

Answer: The probability of individuals receiving either the new page or the old page is approximately equal. Consequently, the results of the conversion rates leads us to conclude that the old page yields more conversions than the new page (12.04% to 11.88%). However, we still need to perform A/B testing to see if this is true or not.

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.

Null: p_{old} **>=** p_{new}

Alternative: p_{old} **<** p_{new}

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 **conversion rate** for p_{new} under the null?

```
In [67]: df2.converted.mean()
```

```
Out[67]: 0.11959708724499628
```

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

```
In [68]: df2.converted.mean()
```

```
Out[68]: 0.11959708724499628
```

c. What is n_{new}, the number of individuals in the treatment group?

```
In [69]: n_new = len(df2t.index)
         n_new
```

```
Out[69]: 145310
```

d. What is n_{old}, the number of individuals in the control group?

```
In [70]: n_old = len(df2c.index)
         n_old
```

```
Out[70]: 145274
```

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

```
In [17]: new_page_converted = np.random.choice([1, 0], size=len(df2t.index), p=[df2.converted.me
```

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

```
In [18]: old_page_converted = np.random.choice([1, 0], size=len(df2c.index), p=[df2.converted.me
```

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

```
In [19]: new_page_converted.mean() - old_page_converted.mean()
```

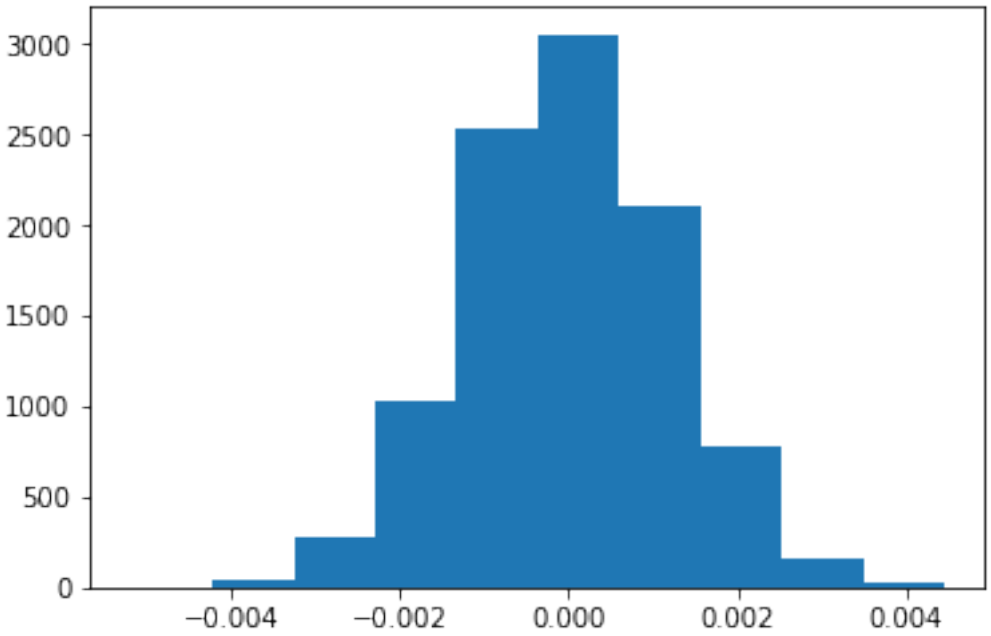
```
Out[19]: -0.0013442939834080458
```

h. Create 10,000 p_{new} - p_{old} values using the same simulation process you used in parts (a) through (g) above. Store all 10,000 values in a NumPy array called **p_diffs**.

```
In [20]: p_diffs = []
         for _ in range(10000):
             new_page_converted = np.random.choice([1, 0], size=len(df2t.index), p=[df2.converte
             old_page_converted = np.random.choice([1, 0], size=len(df2c.index), p=[df2.converte
             p_diffs.append(new_page_converted.mean() - old_page_converted.mean())
```

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



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

```
In [23]: ab_data_diff = df2t.converted.mean() - df2c.converted.mean()
         p_value = (p_diffs > ab_data_diff).mean()
         p_value
```

```
Out[23]: 0.90449999999999997
```

k. Please explain using the vocabulary you've learned in this course 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?

The proportion of the p_diffs (the sampling distribution of the differences of the two separate groups) that is greater than the actual difference is called the p-value

The p-value is the probability of observing the statistic considering that the null hypothesis is true. Since the p-value is 0.9045, we fail to reject the null hypothesis and stick with it.

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 [24]: import statsmodels.api as sm

         convert_old = len(df2c[df2c['converted'] == 1])
         convert_new = len(df2t[df2t['converted'] == 1])
         n_old = len(df2c.index)
         n_new = len(df2t.index)
```

```
/opt/conda/lib/python3.6/site-packages/statsmodels/compat/pandas.py:56: FutureWarning: The panda
  from pandas.core import datetools
```

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

```
Out[27]: (1.3109241984234394, 0.90505831275902449)
```

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.**?

In statistics, the standard score (z-score) is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured.

For a test with CI of 95%, a z-score needed to reject the null hypothesis should not be less than +- 1.96. Since here it is 1.311, we fail to reject the null hypothesis.

Furthermore, the p-value here is 0.9, which leads us to fail to reject the null hypothesis, in agreement with the points j and k.

Part III - A regression approach

1. In this final part, you will see that the result you achieved in the A/B test in Part II above can also be achieved 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.

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 in df2 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 **control**.

```
In [33]: df2['intercept'] = 1
         df2['ab_page']=0
         ab_page_index = df2[df2['group']=='treatment'].index
         df2.loc[ab_page_index, "ab_page"] = 1
```

```
Out[33]:         user_id                   timestamp      group landing_page  \
         43245    646925  2017-01-11 13:15:46.318851  treatment     new_page
         77000    695510  2017-01-03 19:37:30.460998  treatment     new_page
         189734   851049  2017-01-05 17:24:14.877906    control     old_page
         191803   856979  2017-01-12 02:56:09.602189    control     old_page
         24482    769625  2017-01-20 20:13:17.065121  treatment     new_page

                 converted  intercept  ab_page
         43245           0          1        1
         77000           0          1        1
         189734          0          1        0
         191803          0          1        0
         24482           0          1        1
```

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

```
In [35]: log_m = sm.Logit(df2['converted'], df2[['intercept', 'ab_page']])
         results = log_m.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 [37]: results.summary2()
```

```
Out[37]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                  Results: Logit
         =================================================================
         Model:              Logit            No. Iterations:   6.0000
         Dependent Variable: converted        Pseudo R-squared: 0.000
         Date:               2020-12-08 20:26 AIC:              212780.3502
         No. Observations:   290584           BIC:              212801.5095
         Df Model:           1                Log-Likelihood:   -1.0639e+05
         Df Residuals:       290582           LL-Null:          -1.0639e+05
         Converged:          1.0000           Scale:            1.0000
         -----------------------------------------------------------------
                       Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
```

```
------------------------------------------------------------------
intercept    -1.9888    0.0081  -246.6690  0.0000  -2.0046  -1.9730
ab_page      -0.0150    0.0114    -1.3109  0.1899  -0.0374   0.0074
==================================================================

"""
```

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 **Part II**?

The p-value associated with the ab_page is 0.1899.

In part II, the null hypothesis was that the old page had a higher or equal conversion rate to the new page, while the alternative was that the new page would yield a higher conversion rate. (One sided hypothesis test)

In part III, the null hypothesis was that the difference between the conversion rates of the two pages equal zero, while the alternative was that the difference is not equal to zero. (Two sided hypothesis test)

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?

Adding more variables would have its pros and cons.

Pros: Addressing other variables like the time of the day the user signs in might affect the test

Cons: It would add more complexity as some variables might affect others (Collinearity).

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

```
In [38]: df_countries = pd.read_csv('countries.csv')
         country_dummies = pd.get_dummies(df_countries['country'])
         df_new = df_countries.join(country_dummies)
         df3 = df2.set_index('user_id').join(df_new.set_index('user_id'))
         df3 = df2.set_index('user_id').join(df_new.set_index('user_id'))
```

```
Out[38]:                          timestamp      group landing_page  converted  \
         user_id
         661590   2017-01-11 16:55:06.154213  treatment     new_page          0
         853541   2017-01-08 18:28:03.143765  treatment     new_page          0
         679687   2017-01-19 03:26:46.940749  treatment     new_page          1
         817355   2017-01-04 17:58:08.979471  treatment     new_page          1
```

```
          839785   2017-01-15 18:11:06.610965  treatment     new_page          1

                   intercept  ab_page country  CA  UK  US
          user_id
          661590           1        1      US   0   0   1
          853541           1        1      US   0   0   1
          679687           1        1      CA   1   0   0
          817355           1        1      UK   0   1   0
          839785           1        1      CA   1   0   0
```

```
In [39]: Log_m2 = sm.Logit(df3['converted'], df3[['intercept', 'CA', 'UK']])
         results2 = Log_m2.fit()
         results2.summary2()

Optimization terminated successfully.
         Current function value: 0.366116
         Iterations 6


Out[39]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                  Results: Logit
         =================================================================
         Model:              Logit            No. Iterations:   6.0000
         Dependent Variable: converted        Pseudo R-squared: 0.000
         Date:               2020-12-08 20:39 AIC:              212780.8333
         No. Observations:   290584           BIC:              212812.5723
         Df Model:           2                Log-Likelihood:   -1.0639e+05
         Df Residuals:       290581           LL-Null:          -1.0639e+05
         Converged:          1.0000           Scale:            1.0000
         -----------------------------------------------------------------
                       Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
         -----------------------------------------------------------------
         intercept    -1.9967    0.0068  -292.3145  0.0000  -2.0101  -1.9833
         CA           -0.0408    0.0269    -1.5178  0.1291  -0.0935   0.0119
         UK            0.0099    0.0133     0.7458  0.4558  -0.0161   0.0360
         =================================================================

         """
```

Here, I used the US as the baseline country, looking at the coefficients, they are very low; suggesting that the relationship between the country and the conversion rate is weak.

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 [41]: Log_m3 = sm.Logit(df3['converted'], df3[['intercept', 'ab_page', 'CA', 'UK']])
         results = Log_m3.fit()
         results.summary2()

Optimization terminated successfully.
         Current function value: 0.366113
         Iterations 6


Out[41]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                 Results: Logit
         =================================================================
         Model:              Logit            No. Iterations:   6.0000
         Dependent Variable: converted        Pseudo R-squared: 0.000
         Date:               2020-12-08 20:44 AIC:              212781.1253
         No. Observations:   290584           BIC:              212823.4439
         Df Model:           3                Log-Likelihood:   -1.0639e+05
         Df Residuals:       290580           LL-Null:          -1.0639e+05
         Converged:          1.0000           Scale:            1.0000
         -----------------------------------------------------------------
                      Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
         -----------------------------------------------------------------
         intercept    -1.9893    0.0089  -223.7628  0.0000  -2.0067  -1.9718
         ab_page      -0.0149    0.0114    -1.3069  0.1912  -0.0374   0.0075
         CA           -0.0408    0.0269    -1.5161  0.1295  -0.0934   0.0119
         UK            0.0099    0.0133     0.7433  0.4573  -0.0162   0.0359
         =================================================================

         """
```

The p-value for the ab_page stays close to the value before it. Leading us to fail to reject the null hypothesis. Thus, the e-commerce website should stick with the old page.

Finishing Up

Congratulations! You have reached the end of the A/B Test Results project! 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 [42]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Analyze_ab_test_results_notebook.ipynb'])
```

```
Out[42]: 0
```

```
In [ ]:
```