Analyze_ab_test_results_notebook-Kinnaird-Third_Submission

February 21, 2018

0.1 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!

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 [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:

```
In [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.

```
In [3]: df.count()

Out[3]: user_id         294478
        timestamp       294478
        group           294478
        landing_page    294478
        converted       294478
        dtype: int64
```

c. The number of unique users in the dataset.

```
In [4]: df['user_id'].nunique()

Out[4]: 290584
```

d. The proportion of users converted.

```
In [5]: df.converted.mean()

Out[5]: 0.11965919355605512
```

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

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

Out[6]: 3893
```

f. Do any of the rows have missing values?

```
In [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**.

```
In [8]: drop_df = df[((df['group'] == 'treatment') == (df['landing_page'] == 'new_page')) == Fa
        df2 = df.drop(drop_df)
```

```
In [9]: # Double Check all of the correct rows were removed - this should be 0
        df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) == False].sl
```

```
Out[9]: 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 [10]: df2['user_id'].nunique()
```

```
Out[10]: 290584
```

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

```
In [11]: df2[df2.duplicated(['user_id'])]
```

```
Out[11]:       user_id                   timestamp      group landing_page  converted
         2893   773192  2017-01-14 02:55:59.590927  treatment     new_page          0
```

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

```
In [12]: df2[df2.duplicated(['user_id'], keep=False)]
```

```
Out[12]:       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]: df2 = df2.drop_duplicates(['user_id'])
```

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.mean()
```

```
Out[14]: 0.11959708724499628
```

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

```
In [15]: p_old_observed = (df2.query('group == "control"')['converted'] == 1).mean()
         print(p_old_observed)
```

```
0.1203863045
```

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

```
In [16]: p_new_observed = (df2.query('group == "treatment"')['converted'] == 1).mean()
         print(p_new_observed)
```

```
0.118808065515
```

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

```
In [17]: (df2.landing_page == 'new_page').mean()

Out[17]: 0.50006194422266881
```

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.

According to the sample of 290,584 users, approximately 12% of users converted regardless of whether they were in the 'control' or 'treatment' group. Approximated 50% of all landing pages visits were the new page. Based on this even distribution of visits between the new page and the old page and the 12% conversion rate regardless of treatment group, there is currently insufficient evidence to say that the new treatment 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 p_{old} and p_{new}, which are the converted rates for the old and new pages.

H_0: p_{old} p_{new}

H_1: $p_{old} < p_{new}$

`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 **convert rate** for p_{new} under the null?

```
In [18]: p_new = df2.converted.mean()
         print(p_new)
```

```
0.11959708724499628
```

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

```
In [19]: p_old = df2.converted.mean()
         print(p_old)
```

```
0.11959708724499628
```

c. What is n_{new}?

```
In [20]: n_new = len(df2[(df2['landing_page']=='new_page')])
         print(n_new)
```

```
145310
```

d. What is n_{old}?

```
In [21]: n_old = len(df2[(df2['landing_page']=='old_page')])
         print(n_old)
```

```
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 [43]: new_page_converted = np.random.binomial(1, p_new, n_new)
         print(new_page_converted)
```

```
[0 0 0 ..., 0 0 0]
```

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 [44]: old_page_converted = np.random.binomial(1, p_old, n_old)
         print(old_page_converted)
```

```
[0 0 0 ..., 0 1 0]
```

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

```
In [45]: p_new_converted = new_page_converted.mean()
         print(p_new)

         p_old_converted = old_page_converted.mean()
         print(p_old)

         p_null = p_new_converted-p_old_converted
         print(p_null)
```

```
0.11959708724499628
0.11959708724499628
0.000596587505157
```

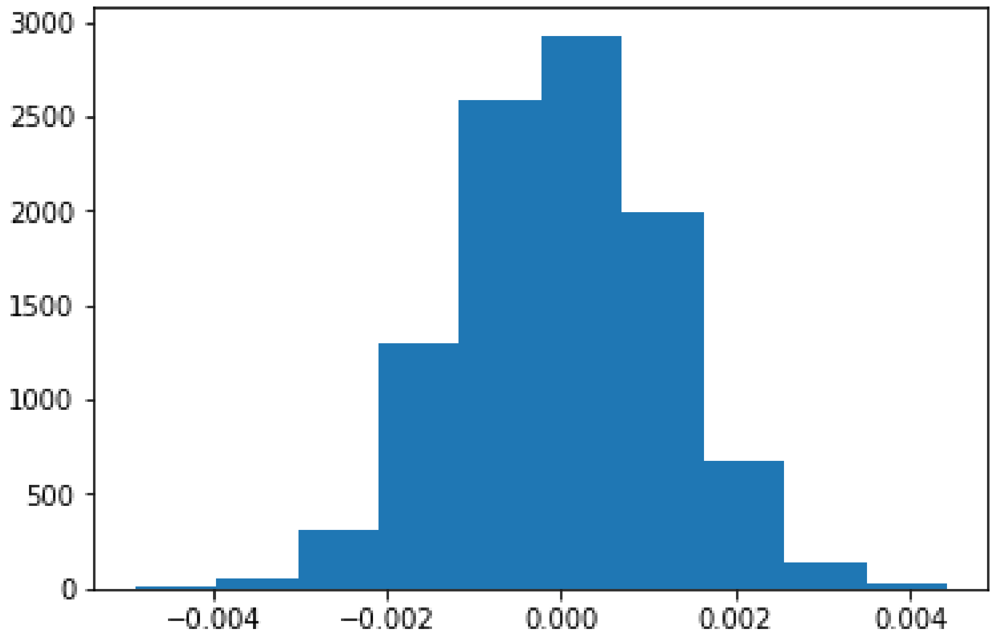
h. Simulate 10,000 p_{new} - p_{old} values using this same process similarly to the one you calculated in parts **a. through g.** above. Store all 10,000 values in a numpy array called **p_diffs**.

```
In [50]: new_converted_simulation = np.random.binomial(n_new, p_new,  10000)/n_new
         old_converted_simulation = np.random.binomial(n_old, p_old,  10000)/n_old
         p_diffs = new_converted_simulation - old_converted_simulation
```

```
In [51]: 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.

```
In [52]: plt.hist(p_diffs);
```



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

```
In [53]: observed_diffs = p_new_observed-p_old_observed
         print(observed_diffs)

         (p_diffs > observed_diffs).mean()

-0.00157823898536


Out[53]: 0.9066999999999995
```

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?

0.9067 is the p-value. Using a type I error rate of 5%, i.e. of 0.05, the null hypothesis is not rejected, because the p-value of 1.0 is greater than the of 0.05.

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 [31]: import statsmodels.api as sm

         convert_old = df2['user_id'][(df2.landing_page == 'old_page') & (df2.converted == 1)]
         convert_new = df2['user_id'][(df2.landing_page == 'new_page') & (df2.converted == 1)]
         n_old = len(df2[(df2['landing_page']=='old_page')])
         n_new = len(df2[(df2['landing_page']=='new_page')])
```

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 [32]: z_score, p_value = sm.stats.proportions_ztest([convert_new, convert_old], [n_new, n_o
         print(z_score, p_value)

-1.31092419842 0.905058312759
```

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 results from the stats.proportions_ztest align with the results from simulated bootstrapped test above. The z-score of -1.3109 and p-value of 0.9051 confirm the null hypothesis should not be rejected, because the z-score is between -1.96 and +1.96 and the p-value is larger than 0.05.

### 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?

Since each row is either a conversion or no conversion, a logistic regression will be used.

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 **control**.

```
In [33]: df2['ab_page'] = np.where(df2['group']=='treatment', 1, 0)
         df2['intercept'] = 1
         df2.head()
```

```
Out[33]:    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
         0        0          1
         1        0          1
         2        1          1
         3        1          1
         4        0          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 [34]: log_mod = sm.Logit(df2['converted'], df2[['intercept', 'ab_page']])
         results = log_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 [35]: results.summary()
```

```
Out[35]: <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:                Wed, 21 Feb 2018   Pseudo R-squ.:               8.077e-06
         Time:                        16:20:29   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 **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**?

According to the Logit Regression Results, the p-value associated with 'ab_page' is 0.190. The value found here, 0.190, is different from the values found from the earlier analyses, 0.9067 and 0.9051. The difference in p-values relates to the difference in null and alternative hypotheses. The hypothese used for the original calculations: H_0**:** p_{old}** p_{new} H_1: $p_{old} < p_{new}$ The hypotheses used for the Logit Regression: H_0: $p_{old} = 0$ H_1: p_{old} 0 **

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?

In addition to the factors already considered (group and landing page), other factors need also be considered. Considering other factors such as user location, time of year of visit, or number of visits per user creates a clearer understanding of users' habits, thus improving the ability to motivate user conversion. While additional factors can offer clarity, abusing the addition of factors will begin to skew results. Furthermore, simple correlation between factors does not mean a causal relationship exists.

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.

```
In [36]: countries_df = pd.read_csv('./countries.csv')
         df_new = countries_df.set_index('user_id').join(df2.set_index('user_id'), how='inner'
         ### Create the necessary dummy variables
         country_dummies = pd.get_dummies(df_new['country'])
         df_new = df_new.join(country_dummies)

In [37]: log_mod = sm.Logit(df_new['converted'], df_new[['intercept', 'CA','UK']])
         results = log_mod.fit()
         results.summary()

Optimization terminated successfully.
         Current function value: 0.366116
         Iterations 6


Out[37]: <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:                Wed, 21 Feb 2018   Pseudo R-squ.:               1.521e-05
         Time:                        16:20:46   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
         CA            -0.0408      0.027     -1.518      0.129      -0.093       0.012
         UK             0.0099      0.013      0.746      0.456      -0.016       0.036
         ==============================================================================
         """
```

The CA p-value is 0.129 and the UK p-value is .456. Using a type I error rate of 5%, i.e. of 0.05, neither of these coefficients are significant because the p-values are greater than .

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 [38]: log_mod = sm.Logit(df_new['converted'], df_new[['intercept', 'ab_page','CA','UK']])
         results = log_mod.fit()
         results.summary()

Optimization terminated successfully.
         Current function value: 0.366113
         Iterations 6


Out[38]: <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:                Wed, 21 Feb 2018   Pseudo R-squ.:               2.323e-05
         Time:                        16:20:52   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
         ab_page       -0.0149      0.011     -1.307      0.191      -0.037       0.007
         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
         ==============================================================================
         """
```

The ab_page p-value is 0.191, the CA p-value is 0.130, and the UK p-value is .457. Using a type I error rate of 5%, i.e. of 0.05, none of these coefficients are significant becuase the p-values are greater than . Therefore, the null hypothesis should not be rejected.

## 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 beginning the next module in the program.