Analyze A/B Test Results

This project will assure you have mastered the subjects covered in the statistics lessons. We have organized the current notebook into the following sections:

Specific programming tasks are marked with a **ToDo** tag.

Introduction

A/B tests are very commonly performed by data analysts and data scientists. 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 webpage,
- Keep the old webpage, or
- Perhaps run the experiment longer to make their decision.

Each **ToDo** task below has an associated quiz present in the classroom. Though the classroom quizzes are **not necessary** to complete the project, they help ensure you are on the right track as you work through the project, and you can feel more confident in your final submission meeting the rubric (https://review.udacity.com/#!/rubrics/1214/view) specification.

> **Tip**: Though it's not a mandate, students can attempt the classroom quizzes to ensure statistical numeric values are calculated correctly in many cases.

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
random.seed(42)
```

ToDo 1.1

Now, read in the `ab_data.csv` data. Store it in `df`. Below is the description of the data, there are a total of 5 columns:

Data columns	Purpose	Valid values
user_id	Unique ID	Int64 values
timestamp	Time stamp when the user visited the webpage	-
group	In the current A/B experiment, the users are categorized into two broad groups. The `control` group users are expected to be served with `old_page` ; and `treatment` group users are matched with the `new_page` . However, **some inaccurate rows** are present in the initial data, such as a `control` group user is matched with a `new_page` .	`['control', 'treatment']`
landing_page	It denotes whether the user visited the old or new webpage.	`['old_page', 'new_page']`
converted	It denotes whether the user decided to pay for the company's product. Here, `1` means yes, the user bought the product.	`[0, 1]`

Use your dataframe to answer the questions in Quiz 1 of the classroom.

> **Tip**: Please save your work regularly.

a. Read in the dataset from the `ab_data.csv` file and take a look at the top few rows here:

In [2]:
```
df = pd.read_csv('ab_data.csv') # Read the dataset
df.head(10) # Show the first 10 rows of the dataset
```

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
5	936923	2017-01-10 15:20:49.083499	control	old_page	0
6	679687	2017-01-19 03:26:46.940749	treatment	new_page	1
7	719014	2017-01-17 01:48:29.539573	control	old_page	0
8	817355	2017-01-04 17:58:08.979471	treatment	new_page	1
9	839785	2017-01-15 18:11:06.610965	treatment	new_page	1

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

In [3]:
```
len(df)
```

Out[3]: 294478

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 == True).mean()
```
Out[5]:
```
0.11965919355605512
```

e. The number of times when the "group" is `treatment` but "landing_page" is not a `new_page`.

In [6]:
```
len(df[((df.group !='treatment') & (df.landing_page == 'new_page') | (df.grou
```
Out[6]:
```
3893
```

f. Do any of the rows have missing values?

In [7]:
```
df.info()
```
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294478 entries, 0 to 294477
Data columns (total 5 columns):
 #   Column        Non-Null Count   Dtype 
---  ------        --------------   ----- 
 0   user_id       294478 non-null  int64 
 1   timestamp     294478 non-null  object
 2   group         294478 non-null  object
 3   landing_page  294478 non-null  object
 4   converted     294478 non-null  int64 
dtypes: int64(2), object(3)
memory usage: 11.2+ MB
```

ToDo 1.2

In a particular row, the **group** and **landing_page** columns should have either of the following acceptable values:

user_id	timestamp	group	landing_page	converted
XXXX	XXXX	control	old_page	X
XXXX	XXXX	treatment	new_page	X

It means, the `control` group users should match with `old_page` ; and `treatment` group users should matched with the `new_page` .

However, for the rows where `treatment` does not match with `new_page` or `control` does not match with `old_page` , we cannot be sure if such rows truly received the new or old wepage.

Use **Quiz 2** in the classroom to figure out how should we handle the rows where the group and landing_page columns don't match?

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]:
```
# Remove the inaccurate rows, and store the result in a new dataframe df2
df2 = df[((df.group == 'treatment') & (df.landing_page == 'new_page') | (df.g
```

In [9]:
```
len(df2)
```

Out[9]: 290585

In [10]:
```
# Double Check all of the incorrect rows were removed from df2 -
# Output of the statement below should be 0
df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) ==
```

Out[10]: 0

ToDo 1.3

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

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

In [11]:
```
df2.nunique()[0]
```

Out[11]: 290584

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

In [12]:
```
df2[df2['user_id'].duplicated()].iloc[:,:1]
```

Out[12]:

	user_id
2893	773192

c. Display the rows for the duplicate **user_id**?

In [13]:
```
df2[df2['user_id'].duplicated()]
```

Out[13]:

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

d. Remove **one** of the rows with a duplicate **user_id**, from the **df2** dataframe.

In [14]:
```
# Remove one of the rows with a duplicate user_id..
# Hint: The dataframe.drop_duplicates() may not work in this case because the
df2.drop(2893, inplace = True)
# Check again if the row with a duplicate user_id is deleted or not
df2[df2.user_id == 773192]
```

```
D:\Anaconda\Anaconda3\lib\site-packages\pandas\core\frame.py:4308: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://p
andas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vi
ew-versus-a-copy)
  return super().drop(
```

Out[14]:

	user_id	timestamp	group	landing_page	converted
1899	773192	2017-01-09 05:37:58.781806	treatment	new_page	0

In [15]:
```
len(df2)
```

Out[15]:
```
290584
```

ToDo 1.4

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?

> **Tip**: The probability you'll compute represents the overall "converted" success rate in the population and you may call it $p_{population}$.

In [16]:
```
P_converted = df2['converted'].mean()
P_converted
```

Out[16]:
```
0.11959708724499628
```

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

In [17]:
```
P_converted_control = df2[(df2.group == 'control')]['converted'].mean()
P_converted_control
```

Out[17]:
```
0.1203863045004612
```

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

In [18]:
```
P_converted_treatment = df2[(df2.group == 'treatment')]['converted'].mean()
P_converted_treatment
```

Out[18]: 0.11880806551510564

> **Tip**: The probabilities you've computed in the points (b). and (c). above can also be treated as conversion rate. Calculate the actual difference (`obs_diff`) between the conversion rates for the two groups. You will need that later.

In [19]:
```
# Calculate the actual difference (obs_diff) between the conversion rates for
act_diff = P_converted_treatment - P_converted_control
act_diff
```

Out[19]: -0.0015782389853555567

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

In [20]:
```
print('The probability that and individual recived the new page is {}'.format
```

```
The probability that and individual recived the new page is 0.5000619442226
688
```

e. Consider your results from parts (a) through (d) above, and explain below whether the new `treatment` group users lead to more conversions.

> There is no sufficient evidence to say that the new treatment page leads to more conversions as the conversion of treatment group is most similar to the conversion of control gorup and the probability that an individual recived the new page is 50%.

> P_converted_control: 0.1204
>
> P_converted_treatment: 0.1188

Part II - A/B Test

Since a timestamp is associated with each event, you could run a hypothesis test continuously as long as you observe the events.

However, then the hard questions would be:

- 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.

ToDo 2.1

For now, consider you need to make the decision just based on all the data provided.

> Recall that you just calculated that the "converted" probability (or rate) for the old page is *slightly* higher than that of the new page (ToDo 1.4.c).

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 be your null and alternative hypotheses (H_0 and H_1)?

You can state your hypothesis in terms of words or in terms of p_{old} and p_{new}, which are the "converted" probability (or rate) for the old and new pages respectively.

> 1. $H_0 : Pnew \leq Pold$
> 2. $H_1 : Pnew > Pold$

ToDo 2.2 - Null Hypothesis H_0 Testing

Under the null hypothesis H_0, assume that p_{new} and p_{old} are equal. Furthermore, assume that p_{new} and p_{old} both are equal to the **converted** success rate in the `df2` data regardless of the page. So, our assumption is:

$$p_{new} = p_{old} = p_{population}$$

In this section, you will:

- Simulate (bootstrap) sample data set for both groups, and compute the "converted" probability p for those samples.
- Use a sample size for each group equal to the ones in the `df2` data.
- Compute the difference in the "converted" probability for the two samples above.
- Perform the sampling distribution for the "difference in the converted probability" between the two simulated-samples over 10,000 iterations; and calculate an estimate.

Use the cells below to provide the necessary parts of this simulation. 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 hypothesis?

In [21]:
```
P_new = df2['converted'].mean()
P_new
```

Out[21]: 0.11959708724499628

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

In [22]:
```
P_old = df2['converted'].mean()
P_old
```

Out[22]: 0.11959708724499628

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

Hint: The treatment group users are shown the new page.

In [23]:
```
N_new = (df2.landing_page == 'new_page').sum()
N_new
```

Out[23]: 145310

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

In [24]:
```
N_old = (df2.landing_page == 'old_page').sum()
N_old
```

Out[24]: 145274

e. Simulate Sample for the `treatment` Group

Simulate n_{new} transactions with a conversion rate of p_{new} under the null hypothesis.

Hint: Use `numpy.random.choice()` method to randomly generate n_{new} number of values.
Store these n_{new} 1's and 0's in the `new_page_converted` numpy array.

In [25]:
```
# Simulate a Sample for the treatment Group
new_page_converted = np.random.choice([0,1], N_new, p = [P_new,1-P_new])
new_page_converted.mean()
```

Out[25]: 0.8804555777303695

f. Simulate Sample for the `control` Group

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

In [26]:
```
# Simulate a Sample for the control Group
old_page_converted = np.random.choice([0,1], N_old, p = [P_new,1-P_new])
old_page_converted.mean()
```

Out[26]: 0.8824290650770269

g. Find the difference in the "converted" probability $(p'_{new} - p'_{old})$ for your simulated samples from the parts (e) and (f) above.

In [27]:
```
new_page_converted.mean() - old_page_converted.mean()
```

Out[27]: -0.0019734873466573655

h. Sampling distribution
Re-create `new_page_converted` and `old_page_converted` and find the $(p'_{new} - p'_{old})$ value 10,000 times using the same simulation process you used in parts (a) through (g) above.

Store all $(p'_{new} - p'_{old})$ values in a NumPy array called `p_diffs`.

In [28]:
```
# Sampling distribution
P_diffs = []
for x in range(10000):
    sim_new_page_converted = np.random.choice([0,1], N_new, p = [P_new, 1-P_r
    sim_old_page_converted = np.random.choice([0,1], N_old, p = [P_old, 1-P_c
    P_diffs.append(sim_new_page_converted.mean() - sim_old_page_converted.mea
```

i. Histogram
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.

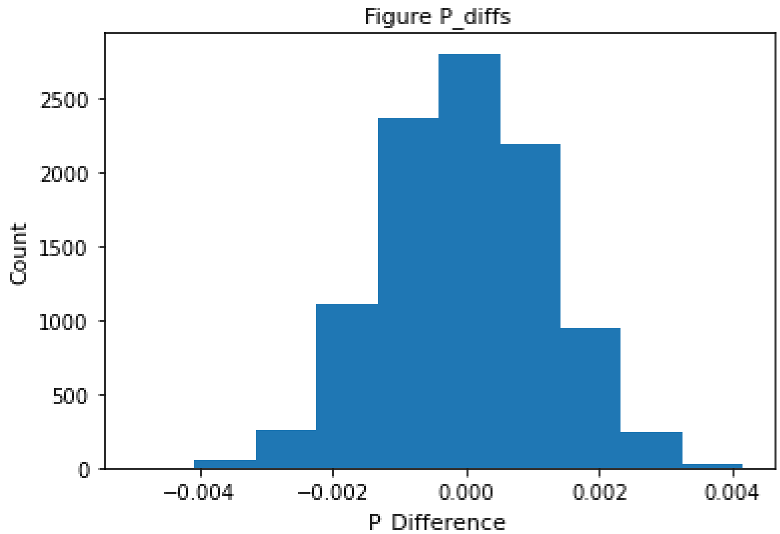
Also, use `plt.axvline()` method to mark the actual difference observed in the `df2` data (recall `obs_diff`), in the chart.

> **Tip**: Display title, x-label, and y-label in the chart.

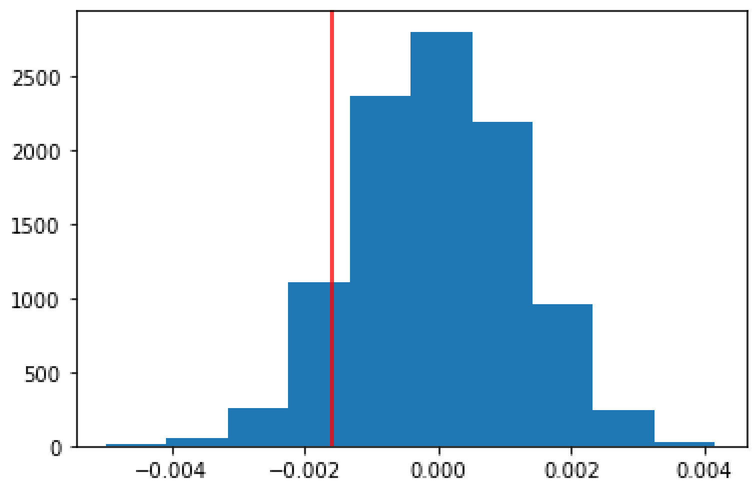
In [29]:
```
P_diffs_arr = np.array(P_diffs)

def hist_figure(x, title, yl, xl):
    plt.hist(x)
    plt.title(title,fontsize=(11))
    plt.ylabel(yl,fontsize=(11))
    plt.xlabel(xl,fontsize=(11))
```

In [30]:
```
hist_figure(P_diffs_arr,'Figure P_diffs','Count','P_Difference')
```



In [31]:
```
plt.hist(P_diffs_arr)
plt.axvline(x=act_diff, color = 'r');
```



j. What proportion of the **p_diffs** are greater than the actual difference observed in the `df2` data?

In [32]:
```
(P_diffs_arr > act_diff).mean()
```

Out[32]: 0.8997

k. Please explain in words what you have just computed in part **j** above.

- What is this value called in scientific studies?
- What does this value signify in terms of whether or not there is a difference between the new and old pages? *Hint*: Compare the value above with the "Type I error rate (0.05)".

It called P_Value.

P Value is greater than 0.05 (Alpha) so we cannot reject the null hypothesis.

So we don't have evidence that the new page is better than old page.

l. Using Built-in Methods for Hypothesis Testing
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 statements below to calculate the:

- `convert_old` : number of conversions with the old_page
- `convert_new` : number of conversions with the new_page
- `n_old` : number of individuals who were shown the old_page
- `n_new` : number of individuals who were shown the new_page

In [33]:
```
import statsmodels.api as sm

# number of conversions with the old_page
convert_old = len(df2[(df2.landing_page == 'old_page') & (df2.converted == 1)
# number of conversions with the new_page
convert_new = len(df2[(df2.landing_page == 'new_page') & (df2.converted == 1)

# number of individuals who were shown the old_page
n_old = len(df2[df2.landing_page == 'old_page'])

# number of individuals who received new_page
n_new = len(df2[df2.landing_page == 'new_page'])

print('\nn_old: {} \nconvert_new: {} '.format(convert_old, convert_new))
print('\nn_old: {} \nconvert_new: {} '.format(n_old, n_new))
```

```
n_old: 17489
convert_new: 17264

n_old: 145274
convert_new: 145310
```

m. Now use `sm.stats.proportions_ztest()` to compute your test statistic and p-value. [Here (https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportions_ztest.html)](https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportions_ztest.html) is a helpful link on using the built in.

The syntax is:

```
proportions_ztest(count_array, nobs_array, alternative='larger')
```

where,

- `count_array` = represents the number of "converted" for each group
- `nobs_array` = represents the total number of observations (rows) in each group
- `alternative` = choose one of the values from `[‘two-sided’, ‘smaller’, ‘larger’]` depending upon two-tailed, left-tailed, or right-tailed respectively.

> **Hint**:
> It's a two-tailed if you defined H_1 as $(p_{new} = p_{old})$.
> It's a left-tailed if you defined H_1 as $(p_{new} < p_{old})$.
> It's a right-tailed if you defined H_1 as $(p_{new} > p_{old})$.

The built-in function above will return the z_score, p_value.

About the two-sample z-test

Recall that you have plotted a distribution `p_diffs` representing the difference in the "converted" probability $(p'_{new} - p'_{old})$ for your two simulated samples 10,000 times.

Another way for comparing the mean of two independent and normal distribution is a **two-sample z-test**. You can perform the Z-test to calculate the Z_score, as shown in the equation below:

$$Z_{score} = \frac{(p'_{new} - p'_{old}) - (p_{new} - p_{old})}{\sqrt{\frac{\sigma^2_{new}}{n_{new}} + \frac{\sigma^2_{old}}{n_{old}}}}$$

where,

- p' is the "converted" success rate in the sample
- p_{new} and p_{old} are the "converted" success rate for the two groups in the population.
- σ_{new} and σ_{new} are the standard deviation for the two groups in the population.
- n_{new} and n_{old} represent the size of the two groups or samples (it's same in our case)

> Z-test is performed when the sample size is large, and the population variance is known. The z-score represents the distance between the two "converted" success rates in terms of the standard error.

Next step is to make a decision to reject or fail to reject the null hypothesis based on comparing these two values:

- Z_{score}
- Z_{α} or $Z_{0.05}$, also known as critical value at 95% confidence interval. $Z_{0.05}$ is 1.645 for one-tailed tests, and 1.960 for two-tailed test. You can determine the Z_{α} from the z-table manually.

Decide if your hypothesis is either a two-tailed, left-tailed, or right-tailed test. Accordingly, reject OR fail to reject the null based on the comparison between Z_{score} and Z_{α}. We determine whether or not the Z_{score} lies in the "rejection region" in the distribution. In other words, a "rejection region" is an interval where the null hypothesis is rejected iff the Z_{score} lies in that region.

> Hint:
> For a right-tailed test, reject null if Z_{score} > Z_{α}.
> For a left-tailed test, reject null if Z_{score} < Z_{α}.

Reference:

- Example 9.1.2 on this page (https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics Sample_Problems/9.01%3A_Comparison_of_Two_Population_Means-_Large_Independent_Samples), courtesy www.stats.libretexts.org (http://www.stats.libretexts.org)

> **Tip**: You don't have to dive deeper into z-test for this exercise. **Try having an overview of what does z-score signify in general.**

In [34]:

```
import statsmodels.api as sm
# ToDo: Complete the sm.stats.proportions_ztest() method arguments
z_score, p_value = sm.stats.proportions_ztest([convert_old, convert_new],[n_c

print('\nz_score: {} \np_value: {} '.format(z_score, p_value))
```

```
z_score: 1.3109241984234394
p_value: 0.9050583127590245
```

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

> **Tip**: Notice whether the p-value is similar to the one computed earlier. Accordingly, can you reject/fail to reject the null hypothesis? It is important to correctly interpret the test statistic and p-value.

> Z-Score computed the deviation from the mean of standard deviagtion, and P_Value is evidence against a null hypothesis

> The calculated P_Value is almost similar to the early calculated P_Value in previous part, and that assure that we cannot reject the null hypothesis

Part III - A regression approach

ToDo 3.1

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 in the `df2` data 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** library to fit the regression model you specified in part **a.** above to see if there is a significant difference in conversion based on the page-type a customer receives. However, you first need to create the following two columns in the `df2` dataframe:

1. `intercept` - It should be `1` in the entire column.
2. `ab_page` - It's a dummy variable column, having a value `1` when an individual receives the **treatment**, otherwise `0`.

In [35]:
```
group_dummies = pd.get_dummies(df2['group'])['treatment']
df3 = df2.join(group_dummies)
df3 = df3.rename(columns = {'treatment':'ab_page'})
df3['intercept'] = 1
df3.head(10)
```

Out[35]:

	user_id	timestamp	group	landing_page	converted	ab_page	intercept
0	851104	2017-01-21 22:11:48.556739	control	old_page	0	0	1
1	804228	2017-01-12 08:01:45.159739	control	old_page	0	0	1
2	661590	2017-01-11 16:55:06.154213	treatment	new_page	0	1	1
3	853541	2017-01-08 18:28:03.143765	treatment	new_page	0	1	1
4	864975	2017-01-21 01:52:26.210827	control	old_page	1	0	1
5	936923	2017-01-10 15:20:49.083499	control	old_page	0	0	1
6	679687	2017-01-19 03:26:46.940749	treatment	new_page	1	1	1
7	719014	2017-01-17 01:48:29.539573	control	old_page	0	0	1
8	817355	2017-01-04 17:58:08.979471	treatment	new_page	1	1	1
9	839785	2017-01-15 18:11:06.610965	treatment	new_page	1	1	1

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

In [36]:
```
lmodel = sm.Logit(df3['converted'], df3[['intercept','ab_page']])
results = lmodel.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.summary()
```

Out[37]:

Logit Regression Results

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290582
Method:	MLE	**Df Model:**	1
Date:	Fri, 22 Apr 2022	**Pseudo R-squ.:**	8.077e-06
Time:	13:05:49	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
Covariance Type:	nonrobust	**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**?

Hints:

- 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**?
- You may comment on if these hypothesis (Part II vs. Part III) are one-sided or two-sided.
- You may also compare the current p-value with the Type I error rate (0.05).

P_Value: 0.190

The difference between p-values in partII and partIII due to we have performed a one-tailed test in partII, and a two-tailed test in partIII.

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?

It is a good idea as it inherently improves the fit with considering the disadvantages that keep adding more factors could make the model worse

g. Adding countries

Now along with testing if the conversion rate changes for different pages, also add an effect based on which country a user lives in.

1. You will need to read in the **countries.csv** dataset and merge together your `df2` datasets on the appropriate rows. You call the resulting dataframe `df_merged`. [Here (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.join.html)](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.join.html) are the docs for joining tables.
2. Does it appear that country had an impact on conversion? To answer this question, consider the three unique values, `['UK', 'US', 'CA']`, in the `country` column. Create dummy variables for these country columns.

> **Hint:** Use `pandas.get_dummies()` to create dummy variables. **You will utilize two columns for the three dummy variables.**

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

In [38]:

```
# Read the countries.csv
countries_df = pd.read_csv('countries.csv')
```

In [39]:

```
# Join with the df3 dataframe
df_new = countries_df.set_index('user_id').join(df3.set_index('user_id'), how
#View first 5 row
df_new.head(10)
```

Out[39]:

user_id	country	timestamp	group	landing_page	converted	ab_page	intercept
834778	UK	2017-01-14 23:08:43.304998	control	old_page	0	0	1
928468	US	2017-01-23 14:44:16.387854	treatment	new_page	0	1	1
822059	UK	2017-01-16 14:04:14.719771	treatment	new_page	1	1	1
711597	UK	2017-01-22 03:14:24.763511	control	old_page	0	0	1
710616	UK	2017-01-16 13:14:44.000513	treatment	new_page	0	1	1
909908	UK	2017-01-06 20:44:26.334764	treatment	new_page	0	1	1
811617	US	2017-01-02 18:42:11.851370	treatment	new_page	1	1	1
938122	US	2017-01-10 09:32:08.222716	treatment	new_page	1	1	1
887018	US	2017-01-06 11:09:40.487196	treatment	new_page	0	1	1
820683	US	2017-01-14 11:52:06.521342	treatment	new_page	0	1	1

In [40]:

```
print(df_new['country'].unique())
```

```
['UK' 'US' 'CA']
```

In [41]:

```
# Create the necessary dummy variables
country_dummies = pd.get_dummies(df_new['country'])
df_new = df_new.join(country_dummies)
df_new.head()
```

Out[41]:

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

In [42]:
```
# Fit your model, and summarize the results
lmodel2 = sm.Logit(df_new['converted'], df_new[['ab_page', 'intercept', 'CA',
results2 = lmodel2.fit()
results2.summary()
```

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

Out[42]:

Logit Regression Results

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290580
Method:	MLE	**Df Model:**	3
Date:	Fri, 22 Apr 2022	**Pseudo R-squ.:**	2.323e-05
Time:	13:05:51	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.1760

	coef	std err	z	P>|z|	[0.025	0.975]
ab_page	-0.0149	0.011	-1.307	0.191	-0.037	0.007
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

> Based on P_Values of lmodel2 that show that there no significant effect on conversation due to all P_Values higher than 0.05 (Alpha)

h. Fit your model and obtain the results
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 are there significant effects on conversion. **Create the necessary additional columns, and fit the new model.**

Provide the summary results (statistical output), and your conclusions (written response) based on the results.

> **Tip**: Conclusions should include both statistical reasoning, and practical reasoning for the situation.

> **Hints**:

- Look at all of p-values in the summary, and compare against the Type I error rate (0.05).
- Can you reject/fail to reject the null hypotheses (regression model)?

- Comment on the effect of page and country to predict the conversion.

In [43]:

```
df_new['UK_ab_page'] = df_new['UK']*df_new['ab_page']
df_new['US_ab_page'] = df_new['US']*df_new['ab_page']
df_new.head()
```

Out[43]:

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

In [44]:

```python
# Fit your model, and summarize the results
lmodel3 = sm.Logit(df_new['converted'], df_new[[ 'intercept','ab_page','UK','
results3 = lmodel3.fit()
results3.summary()
```

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

Out[44]:

Logit Regression Results

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290578
Method:	MLE	**Df Model:**	5
Date:	Fri, 22 Apr 2022	**Pseudo R-squ.:**	3.482e-05
Time:	13:05:54	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.1920

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-2.0040	0.036	-55.008	0.000	-2.075	-1.933
ab_page	-0.0674	0.052	-1.297	0.195	-0.169	0.034
UK	0.0118	0.040	0.296	0.767	-0.066	0.090
US	0.0175	0.038	0.465	0.642	-0.056	0.091
UK_ab_page	0.0783	0.057	1.378	0.168	-0.033	0.190
US_ab_page	0.0469	0.054	0.872	0.383	-0.059	0.152

lmodel3 results summary

The p-value of the interaction (UK_ab_page / US_ab_page) is higher than 0.05.

Adding (UK_ab_page / US_ab_page) to the regression model fails to provide any statistical evidance that there is any impact on the convertion.

Conclusions

Based on the above analysis & conclusions we couldn't reject the null hypothesis, and so we don't have any evidence that the conversion of the new page is higher than the converion of the old page So based on that we recommend that the company stick to the old page as it is better with a very miniscule value.