Mobile App A/B Test

Louis Bailey

Introduction

There is a mobile app with two variations for an enrollment button. The control group says 'Secure Free Trial', and the experimental group says 'Enroll Now'. The goal is to see if changing to the experimental group - 'Enroll Now' will result in more clicks and boost the company's sales.

In this experiment a two-sided two-sample z-test will be used because we do not have a specific hypothesis about which group might perform better, and we are equally interested in deviations in either direction. Specifically this is a a two-sided two-sample z-test for proportions.

In [1]:

```
import pandas as pd
import numpy as np

from scipy.stats import norm
import statsmodels.stats.power as smp

import matplotlib.pyplot as plt
import seaborn as sns
```

$$\begin{cases} H_0 : p_{con} = p_{exp} \\ H_1 : p_{con} \neq p_{exp} \end{cases}$$

Loading Click Rate Data for AB Test

In [2]:

```
df_ab_test = pd.read_csv('https://raw.githubusercontent.com/TatevKaren/CaseStudies/main/AB%20Testing/ab_test_click
```

In [3]:

```
print(df_ab_test.shape)
df_ab_test.head()
```

```
(20000, 4)
```

Out[3]:

	user_id	click	group	timestamp
0	1	1	exp	2024-01-01 00:00:00
1	2	0	exp	2024-01-01 00:01:00
2	3	1	exp	2024-01-01 00:02:00
3	4	0	exp	2024-01-01 00:03:00
4	5	1	exp	2024-01-01 00:04:00

Description of Data

In [4]:

```
df_ab_test['group'].value_counts()
```

Out[4]:

```
group
exp     10000
con     10000
Name: count, dtype: int64
```

- The samples are very large, so for the 2-sided-2-sample z test the assumptions should be met by the cental limit theorem.
- Also, we can see the groups are balanced 50/50.

In [5]:

```
df_ab_test[['group', 'click']].groupby('group').sum('click')
```

Out[5]:

	click
group	
con	1989
exp	6116

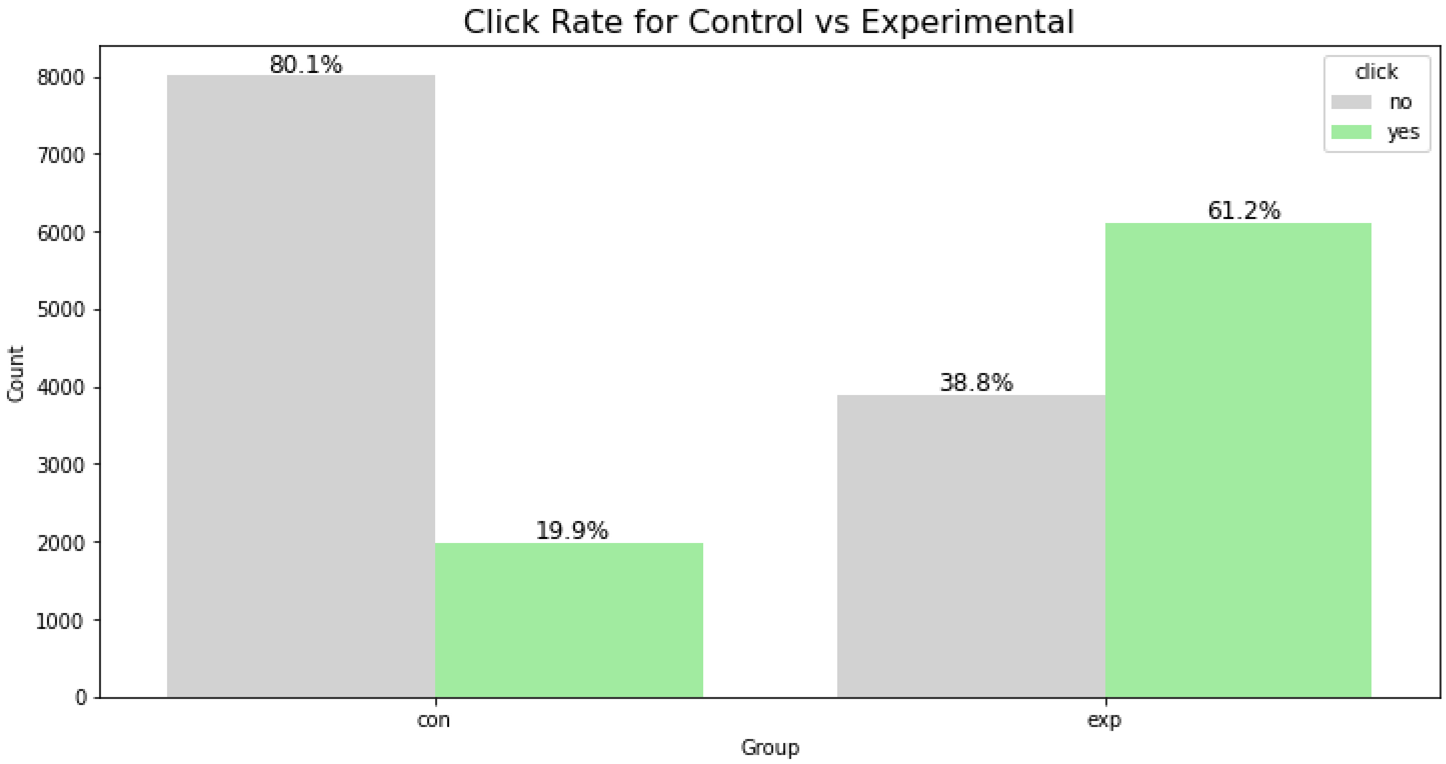
- The experimental group has many more clicks than the control group. We want to know if this difference is statistically and practically significant.

In [6]:

```
df_counts           = df_ab_test.groupby(['click', 'group']).size().reset_index(name='count')
df_counts['click'] = df_counts['click'].map({0: 'no', 1: 'yes'})
palette             = {'no': 'lightgray', 'yes': 'palegreen'}

plt.figure(figsize=(12,6))
sns.barplot(data=df_counts, x='group', y='count', hue='click', palette=palette)
plt.title('Click Rate for Control vs Experimental', size=16)
plt.xlabel('Group')
plt.ylabel('Count')


plt.text(-0.25, 8050, '80.1%', size=12)
plt.text(0.15, 2040, '19.9%', size=12)
plt.text(1.15, 6180, '61.2%', size=12)
plt.text(0.75, 3950, '38.8%', size=12)
plt.show()
```



Parameters of the Power Analysis

- β : probability of type II error
- $(1-\beta)$: power of the test
- α : probability of tye I error, significance level
- δ : Minimum Detectable Effect

In [7]:

```
alpha = 0.05
delta = 0.1
```

- An alpha of 0.05 means the confidence level is 95%

- The minimum detectable effect, delta, is the smallest difference between the control and experimental group that we want to be able to detect with statistical significance. The business in this project has determined that a difference smaller than 10% is not practically significant and does not warrant the effort and resources to implement the change.

- For the power of the test, 1 - beta, we are going with the standard 80%. It is the probability of detecting a true effect or difference between the control and experimental group.

Required Sample Size

In [9]:

```
power = 0.80

analysis = smp.TTestIndPower()
sample_size = analysis.solve_power(effect_size=delta, alpha=alpha, power=power, alternative='two-sided')

print(f"Required sample size per group: {sample_size:.2f}")
```

```
Required sample size per group: 1570.73
```

- The required sample size per group is about 1571. Since we have 10,000 for each group this requirement is met.

Total Number of Clicks Per Group

In [8]:

```
X_con = df_ab_test.groupby('group')['click'].sum().loc['con']
X_exp = df_ab_test.groupby('group')['click'].sum().loc['exp']
```

In [9]:

```
print('Clicks for control:', X_con)
print('Clicks for experimental:', X_exp)
```

```
Clicks for control: 1989
Clicks for experimental: 6116
```

Pooled Estimates

In [10]:

```
#length of each group
N_con = len(df_ab_test[df_ab_test['group']=='con'])
N_exp = len(df_ab_test[df_ab_test['group']=='exp'])
```

In [11]:

```
#estimate of click probabilities per group
p_con_hat = X_con/N_con
p_exp_hat = X_exp/N_exp
```

In [12]:

```
#pooled click probabilities
p_pooled_hat = (X_con + X_exp)/(N_con + N_exp)

print('p_pooled_hat', p_pooled_hat)
```

```
p_pooled_hat 0.40525
```

Pooled Variance

In [13]:

```
pooled_variance = p_pooled_hat * (1-p_pooled_hat) * (1/N_con + 1/N_exp)

print('pooledvariance:',pooled_variance)
```

```
pooledvariance: 4.82044875e-05
```

Standard Error and Test Statistics

In [14]:

```
se        = np.sqrt(pooled_variance)
test_stat = (p_con_hat - p_exp_hat)/se
Z_crit    = norm.ppf(1-alpha/2) #two sided so /2

print('Stadard Error:', se)
print('Test Statistic:', test_stat)
print('Critical Value:', Z_crit)
```

```
Stadard Error: 0.006942945160376826
Test Statistic: -59.44163326469381
Critical Value: 1.959963984540054
```

- The absolute value of the test statistic being much greater than the critical value indcates strong evidence against the null hypothesis.

P Value of the Two Sample Z-Test

In [15]:

```
p_value = 2*norm.sf(abs(test_stat))
print('p value', p_value)

if p_value >= alpha:
    print('Fail to reject null hypothesis')
elif p_value < alpha:
    print('Reject null hypothesis')
```

```
p value 0.0
Reject null hypothesis
```

- Because the p value is less than our chosen alpha of 0.05 we reject the null hypothesis and conclude the experimental group performs significantly different from the control group.

- From the analysis above we can say there is strong evidence the experimental group performs significantly better than the control group.

Normal Distribution with Rejection Region

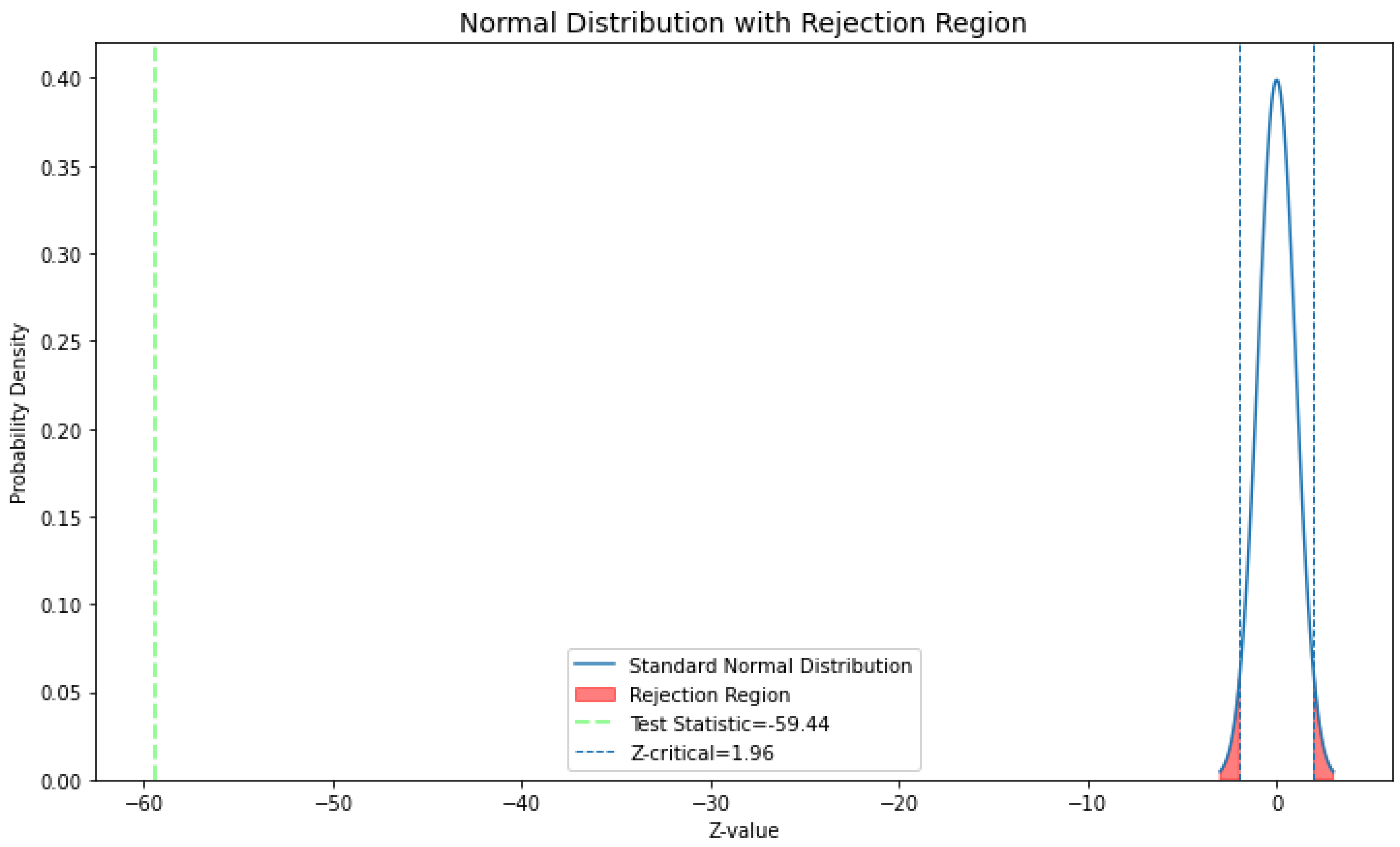
In [26]:

```
mu    = 0
sigma = 1
x     = np.linspace(mu-3*sigma, mu + 3*sigma, 100)
y     = norm.pdf(x, mu, sigma)

plt.figure(figsize=(12,7))

plt.plot(x,y, label='Standard Normal Distribution')
plt.fill_between(x,y, where=(x > Z_crit) | (x < -Z_crit), color='r', alpha=0.5, label='Rejection Region')
plt.axvline(test_stat, color='palegreen', linestyle='--', linewidth=2, label=f'Test Statistic={test_stat:.2f}')
plt.axvline(Z_crit, color='tab:blue', linestyle='--', linewidth=1, label=f'Z-critical={Z_crit:.2f}')
plt.axvline(-Z_crit, color='tab:blue', linestyle='--', linewidth=1)

plt.ylim(0,0.42)
plt.xlabel('Z-value')
plt.ylabel('Probability Density')
plt.title('Normal Distribution with Rejection Region', size=14)
plt.legend()
plt.show()
```



95% Confidence Interval

```
In [27]:  CI = [round((p_exp_hat - p_con_hat) - se*Z_crit,3),
                round((p_exp_hat - p_con_hat) + se*Z_crit,3)]
```

```
In [29]:  CI
```

```
Out[29]: [0.399, 0.426]
```

- The confidence interval being very narrow is a good sign because it indicates high precision of the test- providing a reliable estimate of the true effect.

- The lower bound of the confidence interval means the experimental group has a click rate that is at least 39.9 percentage points higher - well above our delta of 10%

Conclusion

- In conclusion, the A/B test above indicates the experimental group performs significantly better than the control group both statistically and practically. Therefore there is a difference between the control and experimental group, and that difference is large enough for motivation to change the product.