数据分析报告

试验目标：通过此次试验，判断新旧两个页面对用户转化是否有显著区别
衡量指标：点击率

```
import pandas as pd
import scipy as sp
import matplotlib.pyplot as plt
```

```
#加载数据
path = '../huxin/ab_data.csv'
data = pd.read_csv(path, encoding = 'utf-8')
data.head(5)
```

	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

```
#数据清洗
data.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
```

```
data.duplicated().sum()
```

```
0
```

```
data.isnull().sum()
```

```
user_id         0
timestamp       0
group           0
landing_page    0
converted       0
dtype: int64
```

```
#查看流量分配比例，新页面和老页面用户比，比例基本一致
data['group'].value_counts()
```

```
treatment    147276
control      147202
Name: group, dtype: int64
```

```
#检查最小样本量
data[data.landing_page == 'old_page']['converted'].mean()
```

```
0.12047759085568362
```

老页面的点击率为12%，假设我们希望新页面能够让点击率至少提升一个百分点，则算得所需最小样本量为16753。147202>16753满足最小样本量需求。

```
#查看两种页面点击率
plt.rcParams['font.sans-serif'] = ['SimSun']  # 中文字体设置-黑体
```

```
plt.rcParams['axes.unicode_minus'] = False  # 解决保存图像是负号'-'显示为方块的问题
n_old = len(data[data.landing_page == 'old_page'])
n_new = len(data[data.landing_page == 'new_page'])
c_old = len(data[data.landing_page == 'old_page'][data.converted == 1])
c_new = len(data[data.landing_page == 'new_page'][data.converted == 1])
#n_old = len(data[data.landing_page == 'old_page']) #对照组
#n_new = len(data[data.landing_page == 'new_page']) #策略二

try:
    if c_new ==0:
        print('no calculation')
    else:
        r_old = c_old/n_old
        r_new = c_new/n_new
except:
    print("除数为0")
#总和点击率
r = (c_old + c_new) / (n_old + n_new)
print("总和点击率：", r)
print("新版本点击率：", r_new)
print("1版本点击率：", r_old)
#print(c_new,c_old,n_new,n_old)
```

```
总和点击率：  0.12172386392192286
新版本点击率：  0.1229701369881621
1版本点击率：  0.12047759085568362
```

假设检验，假设老页面转化率为p1,新页面转化率为p2
零假设：p1>p2,即p1-p2>0
备择假设：p1<p2,即p1-p2<0
本次实验满足的判断结果只有0和1（转化和未转化），符合0-1分布，独立双样本，总体均值和方差未知，用Z检验

```
#计算检验统计量z
import numpy as np
z = (r_old - r_new) / np.sqrt(r*(1-r)*(1/n_old + 1/n_new))
print("检验统计量z:", z)
```

```
检验统计量z：-2.068408103750818
```

假设a=0.05

```
#看显著水平0.05对应的z的分位数
from scipy.stats import norm
z_alpha = norm.ppf(0.05)
z_alpha
```

```
-1.6448536269514729
```

```
if abs(z) > abs(z_alpha):
    result = "落入拒绝域，拒绝零假设"
else:
    result = "接受零假设"
print(result)
```

```
落入拒绝域，拒绝零假设
```

得出结论：在显著性水平为0.05时，拒绝原假设，新页面转化率更好

```
#求解Cohen’s d系数，衡量效应大小
std_old = data[data.landing_page == "old_page"].converted.std()
std_new = data[data.landing_page == "new_page"].converted.std()
s = np.sqrt(((n_old - 1)* std_old**2 + (n_new - 1)* std_new**2 ) / (n_old +
n_new - 2))
# 效应量Cohen's d
d = (r_old - r_new) / s
print('Cohen\'s d为：', d)
```

```
Cohen's d为：  -0.007623273107908435
```

分析结论

Cohen's d的值约为-0.00762，绝对值很小。两者虽有显著性水平5%时统计意义上的显著差异，但差异的效应量很小。可以理解为显著有差异，但差异的大小不显著。