Домашнее задание 9.3

Задание 3

Сперва подготовим данные, разделим их по вариациям и посчитаем для каждой конверсию в заполнение формы среди уникальных пользователей.

Считаем данные, рисунок 1

```
In [1]:
1 import pandas as pd
2 import numpy as np
```

```
In [5]:
1 #загрузка исходной таблицы
2 df = pd.read_excel('task3.xlsx')
```

```
In [6]:
1 df
```

Out[6]:

	person_id	date	variation	submit_flg
0	1	2023-09-13	test	1
1	2	2023-09-15	test	1
2	3	2023-09-16	test	1
3	4	2023-09-15	test	0
4	5	2023-09-12	test	1
...	...	...	...	...
5433	5404	2023-09-02	control	0
5434	5405	2023-09-06	control	1
5435	5406	2023-09-15	control	1
5436	5407	2023-09-15	control	1
5437	5408	2023-09-06	control	0

5438 rows × 4 columns

Рисунок 1

Разделим данные в две таблицы по столбцу ‘variation’, рисунки 1-2:

In [9]:
```
#Разделим на вариации в две таблицы
df1 = df[df['variation'] == 'test']
df2 = df[df['variation'] == 'control']
```

In [10]:
```
df1
```

Out[10]:

	person_id	date	variation	submit_flg
0	1	2023-09-13	test	1
1	2	2023-09-15	test	1
2	3	2023-09-16	test	1
3	4	2023-09-15	test	0
4	5	2023-09-12	test	1
...	...	...	...	...
5420	5392	2023-09-20	test	1
5421	5393	2023-09-17	test	1
5423	5395	2023-09-02	test	0
5424	5396	2023-09-18	test	1
5429	5401	2023-09-05	test	1

2850 rows × 4 columns

Рисунок 2

In [11]:
```
df2
```

Out[11]:

	person_id	date	variation	submit_flg
5	6	2023-09-02	control	1
6	7	2023-09-10	control	0
8	9	2023-09-16	control	1
9	10	2023-09-20	control	1
11	12	2023-09-20	control	1
...	...	...	...	...
5433	5404	2023-09-02	control	0
5434	5405	2023-09-06	control	1
5435	5406	2023-09-15	control	1
5436	5407	2023-09-15	control	1
5437	5408	2023-09-06	control	0

2588 rows × 4 columns

Рисунок 3

Избавимся от дубликатов, оставив только уникальных пользователей, рисунок 4:

```
In [16]:
#Теперь оставим только уникальных пользователей, удалив строки-дубликаты
df1 = df1.drop_duplicates(subset = 'person_id')
df2 = df2.drop_duplicates(subset = 'person_id')

In [23]:
df1.shape

Out[23]: (2834, 4)

In [24]:
df2.shape

Out[24]: (2574, 4)
```

Рисунок 4

Теперь мы знаем размеры выборок, в случае тестовой это 2834 пользователя, а в контрольном варианте 2574 пользователя. Неравное распределение может быть вызвано множеством причин, например технические, если изменилось что-то внутри компании, из-за чего трэк траффика изменился, также возможен случай, что в тестовом варианте увеличился поток новых пользователей, например сезонность и множество причин. Также случайные факторы, как вариабельность и поведенческие привычки пользователей или случайные колебания, что часто встречается при небольших размерах выборок.

Вычислим конверсию для обоих вариантов, рисунок 5:

```
In [32]:
#Посчитаем для каждого варианта конверсию
conversion_test = df1['submit_flg'].sum() / 2834
conversion_control = df2['submit_flg'].sum() / 2574

In [33]:
print(f"Conversion for test: {conversion_test}")
print(f"Conversion for control: {conversion_control}")

Conversion for test: 0.6983062808750882
Conversion for control: 0.7486402486402487
```

Рисунок 5

Как мы видим конверсия упала на 5%, вероятнее всего наша гипотеза по улучшения не подтвердилась, убедимся окончательно при помощи калькулятора.

Занесём все данные в калькулятор, чтобы выяснить принимать ли нам гипотезу или нет, я использовал два ресурса для принятия решения, рисунки 6-7:

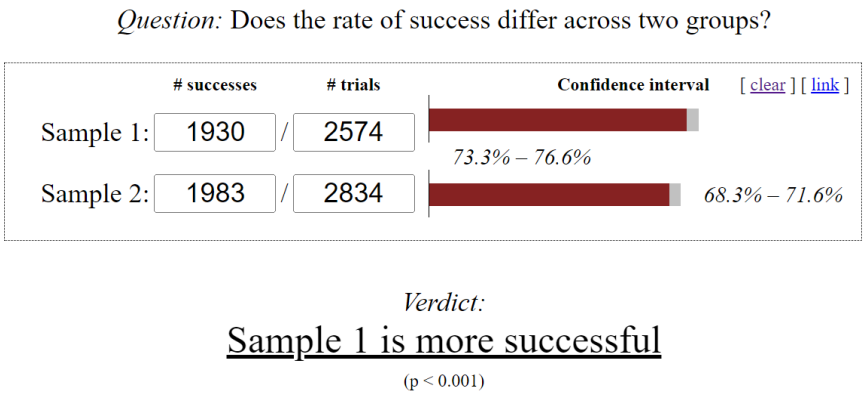

Рисунок 6

Рисунок 7

Подведение итогов:

Конверсия в заполнение формы в тестовой выборке упала на 5% относительно контрольной, этот результат статистически значим и принять гипотезу мы не можем, поэтому тест не нуждается в масштабировании.