Hypothesis Testing with Python · Navigation App

Date: 18 April 2024

Project completed by: Anton Starshev
linkedin.com/in/starshev

Source

This is an end-of-course workplace scenario project **Waze, created in partnership with the realtime driving directions app** proposed within the syllabus of *Google Advanced Data Analytics Professional Certificate* on Coursera.

Purpose

The purpose of this portfolio project is to demonstrate my knowledge of how to prepare, create and conduct hypothesis testing, as well as my ability to draw valuable insights for the benefit of business development.

Context

According to the fictional project scenario, I am working as a data professional in Waze, a free navigation app that makes it easier for drivers around the world to get to where they want to go.

Waze's data team is working on the churn project. An intermediate request from leadership has emerged: **to analyze the relationship between mean amount of rides and device type**.

They require a statistical analysis of ride data based on device type. Specifically, leadership seeks to ascertain if there is a statistically significant difference in the mean number of rides between iPhone® and Android™ users.

Data

This project uses a dataset called **waze_dataset.csv**. It contains synthetic data created for this project in partnership with Waze.

The dataset contains 14,999 rows (each row represents one unique user) and 12 columns.

Project Goal

The practical goal is to apply descriptive statistics and hypothesis testing using Python, analyzing whether there is a relationship between mean amount of rides and device type. For example, to determine whether drivers who use a specific type of device indeed have a higher average number of rides.

Solution

Starting my project, I divided the execution process into four key phases to carry them out step by step:

1. Importing necessary Python packages and loading the dataset
2. Performing Exploratory Data Analysis (EDA) and computing descriptive statistics
3. Conducting hypothesis testing
4. Formulating business insights and recommendations

1 · Data Loading

Imported packages and libraries needed to compute descriptive statistics and conduct a hypothesis test.

In [1]:
```
import pandas as pd
from scipy import stats
```

Loaded the scenario dataset into a DataFrame.

In [6]:
```
df = pd.read_csv("waze_dataset.csv", index_col = 0)
```

2 · Data Exploration

Previewed the loaded data.

In [7]:
```
df.head()
```

Out[7]:

ID	label	sessions	drives	total_sessions	n_days_after_onboarding	total_navigations_fav1	total_naviga
0	retained	283	226	296.748273	2276	208	
1	retained	133	107	326.896596	1225	19	
2	retained	114	95	135.522926	2651	0	
3	retained	49	40	67.589221	15	322	
4	retained	84	68	168.247020	1562	166	

Checked the data size.

In [8]:
```
df.shape
```

Out[8]:
```
(14999, 12)
```

Verified the data types and names of columns.

In [9]:
```
df.dtypes
```

Out[9]:
```
label                     object
sessions                   int64
drives                     int64
total_sessions           float64
n_days_after_onboarding    int64
total_navigations_fav1     int64
total_navigations_fav2     int64
driven_km_drives         float64
duration_minutes_drives  float64
activity_days              int64
driving_days               int64
device                    object
dtype: object
```

Used descriptive statistics to conduct Exploratory Data Analysis (EDA) on the rides data.

In [12]:
```
df[['drives']].describe(include = 'all')
```

Out[12]:

	drives
count	14999.000000
mean	67.281152
std	65.913872
min	0.000000
25%	20.000000
50%	48.000000
75%	93.000000
max	596.000000

Exploring the relationship between device type and the number of rides customers take, one approach within the EDA was to examine the average ride count for each device type.

In [13]:
```
df.groupby('device')[['drives']].mean()
```

Out[13]:

device	drives
Android	66.231838
iPhone	67.859078

Observation: First, I have found that there are just two device categories.

Second, based on my preliminary research analysis, iPhone users tend to take more rides on average than those who use Android. However, this difference could be due to sample variability. So, the next step was to check the statistical significance of this difference through hypothesis testing.

3 · Hypothesis Test

Since one of the variables is categorical, as a first step, I mapped the device category into numerical values, assigning "1" to iPhone devices and "2" to Android devices. Additionally, I added a corresponding column "device_type" to the DataFrame for testing purposes.

In [19]:
```
device_map = {'Android' : 2, 'iPhone' : 1}

df['device_type'] = df.device.map(device_map)

df[['device','device_type']].head()
```

Out[19]:

ID	device	device_type
0	Android	2
1	iPhone	1
2	Android	2
3	iPhone	1
4	Android	2

Stated the null hypothesis and the alternative hypothesis:

H_0: There is no difference in the average number of rides between clients who use iPhones and those who use Android devices.

H_1: There is a difference in the average number of rides between clients who use iPhones and those who use Android devices.

Assigned a **5% significance level** to the hypothesis test.

Determined the type of hypothesis testing: **two-sample two-tailed t-test**.

Filtered the data into two groups based on the device type: iPhone or Android.

In [15]:
```
iphone_drives = df[df['device_type'] == 1].drives

android_drives = df[df['device_type'] == 2].drives
```

Conducted the hypothesis test using SciPy Stats.

In [20]:
```
stats.ttest_ind(a = iphone_drives, b = android_drives, equal_var = False,
alternative = 'two-sided')
```

Out[20]:
```
TtestResult(statistic=1.463523206885235, pvalue=0.143351972680206, df=11345.066049381952)
```

Test Result: Given that the p-value of 14.3% is notably higher than the 5% significance level, I failed to reject the null hypothesis.

4 · Insight and Recommendation

Business Insight: Based on the conducted test, the key business insight is that there is no statistically significant difference in the average number of rides between clients who use iPhones and those who use Android devices.

Business Recommendation: *Since the test result revealed no direct correlation between user engagement with the service and the type of device they use, I would recommend exploring various other factors within the context of churn research that may influence the user's ride count and conducting hypothesis tests on them.*

Skills

Throughout this project, I showcased the following professional competencies:

- Working with DataFrame and conducting exploratory data analysis using the Pandas library
- Converting a categorical variable into numerical using the Pandas library
- Preparing and conducting hypothesis testing using the SciPy Stats library
- Evaluating test results and formulating data-driven recommendations

Acknowledgment

I would like to express gratitude to Google and Coursera for supporting the educational process and providing the opportunity to refine and showcase skills acquired during the courses by completing real-life scenario portfolio projects, such as this.