Hypothesis Testing with Python · Short Videos App

Date: 18 April 2024

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

Source

This is an end-of-course workplace scenario project **TikTok, created in partnership with the short-form video hosting company** proposed within the syllabus of *Google Advanced Data Analytics Professional Certificate* on Coursera.

Purpose

The objective of this task is to demonstrate my proficiency in utilizing statistical methods for data analysis and interpretation. This includes employing descriptive statistics and hypothesis testing techniques. Additionally, it aims to showcase my ability to organize and communicate crucial information effectively, thereby contributing to business development efforts.

Context

According to the fictional project scenario, I am a member of TikTok's data analytics team that has completed the first three milestones of the claims classification project.

Project management officers inform the data team about a new request: **to determine whether there is a statistically significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts**.

A final email from the Data Scientist details my next assignment: to conduct a hypothesis test on verified versus unverified accounts in terms of video view count.

Data

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

The dataset contains 19,382 rows (each row represents a different published TikTok video in which a claim / opinion has been made) and 11 columns.

Project Goal

The practical objective is to utilize descriptive and inferential statistics, probability distributions and hypothesis testing in Python to analyze whether a genuine relationship exists between user account type and video view count.

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. Wrangling and exploring the project data
3. Implement a hypothesis test
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 [5]:
```
df = pd.read_csv("tiktok_dataset.csv", index_col = 0)
```

2 · Data Wrangling and Exploration

Previewed the loaded data.

In [20]:
```
df.head(3)
```

Out[20]:

#	claim_status	video_id	video_duration_sec	video_transcription_text	verified_status	author_ban_sta
1	claim	7017666017	59	someone shared with me that drone deliveries a...	not verified	under rev
2	claim	4014381136	32	someone shared with me that there are more mic...	not verified	ac
3	claim	9859838091	31	someone shared with me that american industria...	not verified	ac

Checked the data size.

In [7]:
```
df.shape
```

Out[7]:
```
(19382, 11)
```

Verified the data types and names of columns.

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

Out[8]:
```
claim_status                 object
video_id                      int64
video_duration_sec            int64
video_transcription_text     object
verified_status              object
author_ban_status            object
video_view_count            float64
video_like_count            float64
video_share_count           float64
video_download_count        float64
video_comment_count         float64
dtype: object
```

Used descriptive statistics to conduct Exploratory Data Analysis (EDA) on the video view counts.

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

Out[12]:

	video_view_count
count	19084.000000
mean	254708.558688
std	322893.280814
min	20.000000
25%	4942.500000
50%	9954.500000
75%	504327.000000
max	999817.000000

Checked for and handled missing values.

In [14]:
```
df.isnull().sum()
```

Out[14]:
```
claim_status                298
video_id                      0
video_duration_sec            0
video_transcription_text    298
verified_status               0
author_ban_status             0
video_view_count            298
video_like_count            298
video_share_count           298
video_download_count        298
video_comment_count         298
dtype: int64
```

In [21]:
```
df.dropna(axis = 0, inplace = True)
df.isnull().sum()
```

Out[21]:
```
claim_status                0
video_id                    0
video_duration_sec          0
video_transcription_text    0
verified_status             0
author_ban_status           0
video_view_count            0
video_like_count            0
video_share_count           0
video_download_count        0
video_comment_count         0
dtype: int64
```

Checked for duplicated rows.

In [16]:
```
df.duplicated().sum()
```

Out[16]:
```
0
```

Since I was interested in the relationship between account status and video view count, one approach was to examine the mean value of video view count for each group of verified or not verified accounts in the sample data.

In [17]:
```
df.groupby('verified_status')[['video_view_count']].mean()
```

Out[17]:

verified_status	video_view_count
not verified	265663.785339
verified	91439.164167

Observation: Firstly, it is now confirmed that we have only two account status groups for research.

Secondly, according to my initial exploration, videos posted by verified users tend to receive significantly fewer views on average than those published by unverified accounts. However, I needed to demonstrate that this disparity was not a result of sample variability. Therefore, the subsequent step was to assess the statistical significance of this distinction through hypothesis testing.

3 · Hypothesis Test

Stated the null hypothesis and the alternative hypothesis:

H_0: There is no distinction in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts (any divergence observed in the sample data is attributable to chance or sampling variability).

H_1: There is a difference in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts (any observed difference in the sample data is due to an actual difference in the corresponding population means).

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 account status: verified or not verified.

In [18]:
```
verified = df[df['verified_status'] == 'verified'].video_view_count

not_verified = df[df['verified_status'] == 'not verified'].video_view_count
```

Conducted the hypothesis test using SciPy Stats.

In [19]:
```
stats.ttest_ind(a = verified, b = not_verified, equal_var = False,
alternative = 'two-sided')
```

Out[19]:
```
TtestResult(statistic=-25.499441780633777, pvalue=2.6088823687177823e-120, df=1571.16307
4387424)
```

Test Result: Since the p-value is significantly lower than the 5% significance level, I rejected the null hypothesis.

4 · Insight and Recommendation

Business Insight: Based on the conducted test, the key business insight is that there is a statistically significant difference in the average number of views of videos created by verified versus unverified accounts. Specifically, unverified accounts receive much more attention.

Communication and Recommendation: *The analysis revealed potential fundamental behavioral disparities between the account categories. Exploring the underlying cause of this behavioral contrast would be intriguing. For instance, do unverified accounts tend to share more engaging videos, or are unverified accounts associated with any kind of spam bots.*

Skills

Throughout this project, I showcased the following professional competencies:

- Data wrangling and conducting exploratory analysis 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.