TikTok Video Views Hypothesis Testing

August 5, 2023

1 TikTok Video Views Hypothesis Testing

1.0.1 Task 1. Imports and Data Loading

[1]:

```
# Import packages for data manipulation
import pandas as pd
import numpy as np

# Import packages for data visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Import packages for statistical analysis/hypothesis testing
from scipy import stats
```

[2]:

```
# Load dataset into dataframe
data = pd.read_csv("tiktok_dataset.csv")
```

1.0.2 Task 2. Data exploration

[3]:

```
# Display first few rows
data.head()
```

[3]:

```
   # claim_status    video_id  video_duration_sec  \
0  1        claim  7017666017                  59
1  2        claim  4014381136                  32
2  3        claim  9859838091                  31
3  4        claim  1866847991                  25
4  5        claim  7105231098                  19

                            video_transcription_text verified_status  \
0  someone shared with me that drone deliveries a…    not verified
1  someone shared with me that there are more mic…    not verified
2  someone shared with me that american industria…    not verified
3  someone shared with me that the metro of st. p…    not verified
```

```
4  someone shared with me that the number of busi…    not verified

  author_ban_status  video_view_count  video_like_count  video_share_count  \
0      under review          343296.0           19425.0              241.0
1            active          140877.0           77355.0            19034.0
2            active          902185.0           97690.0             2858.0
3            active          437506.0          239954.0            34812.0
4            active           56167.0           34987.0             4110.0

   video_download_count  video_comment_count
0                   1.0                  0.0
1                1161.0                684.0
2                 833.0                329.0
3                1234.0                584.0
4                 547.0                152.0
```

[4]:
```
# Generate a table of descriptive statistics about the data
data.describe()
```

[4]:
```
                  #      video_id  video_duration_sec  video_view_count  \
count  19382.000000  1.938200e+04        19382.000000      19084.000000
mean    9691.500000  5.627454e+09           32.421732     254708.558688
std     5595.245794  2.536440e+09           16.229967     322893.280814
min        1.000000  1.234959e+09            5.000000         20.000000
25%     4846.250000  3.430417e+09           18.000000       4942.500000
50%     9691.500000  5.618664e+09           32.000000       9954.500000
75%    14536.750000  7.843960e+09           47.000000     504327.000000
max    19382.000000  9.999873e+09           60.000000     999817.000000

       video_like_count  video_share_count  video_download_count  \
count      19084.000000       19084.000000          19084.000000
mean       84304.636030       16735.248323           1049.429627
std       133420.546814       32036.174350           2004.299894
min            0.000000           0.000000              0.000000
25%          810.750000         115.000000              7.000000
50%         3403.500000         717.000000             46.000000
75%       125020.000000       18222.000000           1156.250000
max       657830.000000      256130.000000          14994.000000

       video_comment_count
count         19084.000000
mean            349.312146
std             799.638865
min               0.000000
25%               1.000000
50%               9.000000
75%             292.000000
```

```
max            9599.000000
```

Check for and handle missing values.

[5]:

```
# Check for missing values
data.isna().sum()
```

[5]:

```
#                             0
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
```

[6]:

```
# Drop rows with missing values
data = data.dropna(axis=0)
```

[7]:

```
# Display first few rows after handling missing values
data.head()
```

[7]:

```
   # claim_status    video_id  video_duration_sec  \
0  1        claim  7017666017                  59
1  2        claim  4014381136                  32
2  3        claim  9859838091                  31
3  4        claim  1866847991                  25
4  5        claim  7105231098                  19

                            video_transcription_text verified_status  \
0  someone shared with me that drone deliveries a…    not verified
1  someone shared with me that there are more mic…    not verified
2  someone shared with me that american industria…    not verified
3  someone shared with me that the metro of st. p…    not verified
4  someone shared with me that the number of busi…    not verified

  author_ban_status  video_view_count  video_like_count  video_share_count  \
0      under review          343296.0           19425.0              241.0
1            active          140877.0           77355.0            19034.0
2            active          902185.0           97690.0             2858.0
3            active          437506.0          239954.0            34812.0
4            active           56167.0           34987.0             4110.0
```

```
   video_download_count  video_comment_count
0                   1.0                  0.0
1                1161.0                684.0
2                 833.0                329.0
3                1234.0                584.0
4                 547.0                152.0
```

[8]:

```
# Compute the mean `video_view_count` for each group in `verified_status`
data.groupby("verified_status")["video_view_count"].mean()
```

[8]:

```
verified_status
not verified    265663.785339
verified         91439.164167
Name: video_view_count, dtype: float64
```

1.0.3 Task 3. Hypothesis testing

The null hypothesis: there is no distinction in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts. Any differences observed in the sample data are attributed to random chance or sampling variability.

The alternative hypothesis: there is indeed a difference in the number of views between the two types of accounts, and any differences observed in the sample data are due to an actual disparity in the means of the corresponding populations.

Set the significance level: For this analysis, the significance level is 5%

Determine the appropriate test procedure: Here, we are comparing the sample means between two independent samples. Therefore, we will apply a two-sample -test.

[9]:

```
# Conduct a two-sample t-test to compare means
# Save each sample in a variable
unverified = data[data["verified_status"] == "not verified"]["video_view_count"]
verified = data[data["verified_status"] == "verified"]["video_view_count"]

# Implement a t-test using the two samples
stats.ttest_ind(a=unverified, b=verified, equal_var=False)
```

[9]:

```
Ttest_indResult(statistic=25.499441780633777, pvalue=2.6088823687177823e-120)
```

1.1 Step 4: Conclusions

- Based on the analysis, there is a statistically significant difference in the average view counts between videos from verified accounts and videos from unverified accounts on TikTok.
- This finding indicates the presence of behavioral distinctions between these two account types, which require further investigation.

- The next step in this project would be to build a regression model on the “verified_status” variable to predict user behavior in this group of verified users.