**DS5110 Homework 3 – AB Testing - Yueheng Yuan**

**Part 1. Getting to know your data**

1. *What data is in file "t1\_users\_active\_mins.csv"?*

This table contains active minutes data logged after the experiment started, recording the time users spent on the platform on specific dates.

There are three columns: uid (user ID), dt (date when the active minutes were recorded), and active\_mins (time users spent on the platform for that date).

1. *What data is in file "t2\_users\_variant.csv"?*

This table contains user treatment assignments for the experiment, indicating whether a user was in the control or treatment group.

There are four columns: uid (user ID), variant\_number (experiment variant – 0 for control and 1 for treatment), dt (date when users entered the experiment, which is consistent for all in this case), and signup\_date (the date when a user signed up to the platform).

1. *What data is in file "t3\_users\_active\_mins\_pre.csv"?*

This table contains user activity before the experiment by recording the active time users spent on social media on specific dates.

Three columns are the same as those in "t1\_users\_active\_mins.csv".

1. *What data is in file "t4\_users\_attributes.csv"?*

This table contains user attributes including gender and user types.

There are three columns: uid (user ID), gender, and user\_type (new\_user, non\_reader, reader, or contributor).

1. *What data is in file "table\_schema.txt"?*

This file describes the schema of all four csv files above, including the meaning of each column and the context of the data.

**Part 2. Organizing the Data**

1. *What is the overall objective of this study?*

The objective is to analyze the impact of installing a new feature on user engagement, which is measured by active minutes spent on the platform. In this experiment, users are divided into two subsets by experiencing new features or remaining the same for control.

1. *What data do we need to reach that objective?*

We’d need uid (track individual users), variant\_number (identify control vs treatment groups), dt (login date), and active\_mins (measure user engagement for each login).

1. *How is the data in t1 currently organized?*

The t1 file contains active minutes data after the experiment organized by uid and dt, which control vs treatment groups cannot be identified

1. *How should the data in t1 be organized to be useful?*

To analyze engagement differences between the control and treatment groups, `t1\_user\_active\_min.csv` and `t2\_user\_variant.csv` should be merged using `uid` as foreign keys to assign treatment/control labels. The ideal dataset structure after getting merged consists of four columns: uid, dt, variant\_number, and active\_mins.

1. *Organize it*

The left-join operation is performed by matching values from the t2 file (variant\_number) and added into the t1 data based on uid. The organized dataset is saved under `data/merged\_user\_data\_after\_exp.csv` for further statistical analysis.

**Part 3. Statistical Analysis**

Assume control group as group 1 and treatment as group 2, we may perform the statistical analysis under two scenarios: per login (dt) or per user (uid, accumulating the total active time per user).

*Scenario #1: If we analyze the user engagement per login,*

1. *Is there a statically difference between group 1 and group 2?*

Null Hypothesis (H₀): The new feature doesn’t impact the total time spent on the platform. Any difference observed is due to random chance.

Alternative Hypothesis (H₁): The new feature has impacts on the total time spent on the platform.

For the independent t-test, assuming the data in each group approximately follows a normal distribution, T-value = -1.4674, P-value = 0.1423 > 0.05

Therefore, we failed to reject the null hypothesis, and no significant difference was found between the control and treatment groups.

1. *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per login: 35.34, Median active minutes per login: 5.00

Group 2 (Treatment):

Mean active minutes per login: 40.24, Median active minutes per login: 7.00

The treatment mean is slightly higher than the control group, but the median difference is relatively small. However, the large gap between the mean and median for both groups suggests a possible skewed distribution, which may violate the t-test assumption of normality.

1. *What can you conclude based on that data?*

The statistical test indicates that there is no significant difference in user engagement between the control and treatment groups (p > 0.05). While the treatment group has a slightly higher mean and median active minutes, this difference is not statistically significant, and the observed increase may be due to random variation rather than the new feature installed.

*Scenario #2: If we analyze the user engagement per user,*

1. *Is there a statically difference between group 1 and group 2?*

Null Hypothesis and alternative Hypothesis will be the same.

For the independent t-test, assuming the data in each group approximately follows a normal distribution, T-value = 0.4056, P-value = 0.6850 > 0.05

Again, we failed to reject the null hypothesis, and no significant difference was found between the control and treatment groups.

1. *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per user: 837.64, Median active minutes per user: 52.00

Group 2 (Treatment):

Mean active minutes per user: 784.20, Median active minutes per user: 71.00

Like the first scenario, the treatment mean is slightly higher, and a large gap between the mean and median is observed for both groups, suggesting a possible skewed distribution and may violate the t-test assumption of normality.

1. *What can you conclude based on that data?*

Similarly, though the treatment group has a slightly higher mean and median active minutes, the difference in user engagement between the control and treatment groups is not statistically different (p > 0.05).

**Part 4. Digging a Little Deeper**

1. *Can you trust the results? Why or why not?*

The statistical result is less reliable due to the skewed data.

As mentioned in the second question of part 3, there presents large gaps between the mean and median, suggesting that the data in both groups is likely skewed rather than normally distributed. It indicates that the t-test assumption of normality is violated, and both groups may contain outliers or skewed data.

1. *Is the data normally distributed?*

No, the large gap between the mean and median for both groups indicates that they are skewed.

1. *Plot a box plot of group 1 and group 2.*

*Scenario #1: If we analyze the user engagement per login,*

A white rectangular object with black text

AI-generated content may be incorrect.

*Scenario #2: If we analyze the user engagement per user,*

A white background with black text

AI-generated content may be incorrect.

1. *Are there any outliers?*

Both boxplots observe clear outliers for both control and treatment groups.

1. *What might be causing those outliers? (Hint, look at the data in t1. What is the maximum time a user should possibly have?).*

Both groups consist of outliers such that the maximum recorded active time is 99999.0, while there are only 1440 minutes in a day.

1. *Remove any data point that might be causing outliers.*
2. *Redo part 2 and 3 with the new data without those data points.*

Again, the statistical analysis is performed under two scenarios: per login (dt) or per user (uid, accumulating the total active time per user).

*Scenario #1: If we analyze the user engagement per login,*

* *Is there a statically difference between group 1 and group 2?*

For the independent t-test, T-value = -30.6868, P-value = 0.0000 < 0.05

Therefore, there is a statistically significant difference between the control and treatment groups to reject the null hypothesis.

* *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per login: 19.34, Median active minutes per login: 5.00

Group 2 (Treatment):

Mean active minutes per login: 23.53, Median active minutes per login: 7.00

The median is much lower than the mean in both groups, indicating that the data is right-skewed, which may imply that some users have very high engagement and therefore increases the mean value.

*Scenario #2: If we analyze the user engagement per user,*

* *Is there a statically difference between group 1 and group 2?*

For the independent t-test, assuming the data in each group approximately follows a normal distribution, T-value = -0.0093, P-value = 0.9926 > 0.05

In this case, we failed to reject the null hypothesis, and no significant difference was found between the control and treatment groups. When aggregating active time for each user, the treatment doesn’t show a significant impact in comparison to the control group.

* *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per user: 458.22, Median active minutes per user: 52.00

Group 2 (Treatment):

Mean active minutes per user: 458.40, Median active minutes per user: 71.00

When summing the total active time for each user, the mean is nearly identical. The median for the treatment is slightly higher, but this difference is not statistically significant, which suggests that while individual logins may be longer or more frequent in the treatment group, the impact isn’t significant enough when summed across all logins for a user.

* *What can you conclude based on that data?*

According to the output from both scenarios, the treatment seemed to have an effect on engagement per login, but not on total active time per user. This could imply that the treatment led to more engagement per login, but it didn’t significantly affect the total active time for each user.

1. *What is the new conclusion based on the new data?*

As mentioned above, the new feature installed tended to result in more engagement per login, but it didn’t significantly affect the total active time for each user.

At the login level regardless of user, the treatment had a significant positive effect on user engagement. When looking at total active time per user, there is no statistically significant difference between the control and treatment groups.

**Part 5. Digging Even Deeper**

1. *Why do we care about the data from t3?*

The data for user active time before the experiment can provide a baseline for comparison, which helps assess whether the treatment had a significant impact on engagement. The absence of this baseline may bring uncertainty such that any observed change in active time was caused by the treatment or part of normal user behavior.

1. *Accounting for the data from t3 rerun part 2 and 3.*
2. *Are there any new conclusions?*

**Part 6. Exploring other conclusions**

Can you come up with any other conclusion with the data given in t4? If so, what are they? This is open ended. This is left open ended to allow you to further explore the data that is given.

**Part 7. Summarize Your Results**

Write a summary for each part of this assignment and how it impacted your results