Capstone Project - Inferential Statistics

In this report we will explain how we have applied inferential statistic techniques to the KKbox data. In particular, we have investigated the following questions:

1. Are songs listened from 'My library' tab replay more than songs listened from other tabs?
2. What is the probability of a song to be replayed depending on the tab it was listened from? Probabilities will be expressed as confidence intervals. Are there statistically significant differences?
3. Are songs which belong to less genres replay more than songs which belong to more genres?
4. Is there a significant difference between males and females in the probability of replaying songs?
5. Do younger users replay songs more often than older users?

For more details about how the questions were solved, please check the Jupyter notebook file interferal_analysis.ipynb.

Question 1: Are songs listened from 'My library' tab replay more than songs listened from other tabs?

Based on the initial exploration of the data, it looks like songs listened from 'My library' tab have a higher probability to be listened again. We did a hypothesis test to confirm this result.

- The null hypothesis: the replay probability from 'My library' tab and the rest of the tabs is the same.
- The alternate hypothesis: the replay probability from 'My library' tab is higher than from the rest of the tabs.

H0: pl=po versus Ha: pl≠po

Where p is the probability of a user to replay a song.

We used as test statistic the difference in replay probabilities between both samples.

In order to perform our hypothesis test we did the following steps:

- Extract data target variable for ‘my library’ entries and the rest of the tabs entries to create our samples. Note: target variable equals 1 means there are recurring listening event(s) triggered within a month after the user’s very first observable listening event, otherwise target equals 0.
- Compute bootstrap replicates of the Bernoulli probability for each sample.
- Compute the difference of bootstrap replicates between both samples.
- Compute p-value: % of times where the difference of bootstrap replicates is equal or higher than zero

We obtained a p-value of 1. Since the p-value is really high, we reject the null hypothesis in favor of the alternative one. The minimum difference observed in the replicates between 'My library' tab probability and the others is around 23.2%.

We conclude that the probability of replaying a song when it is listened from 'My library' is higher than when it is listened from other tab.

Question 2: What is the probability of a song to be replayed depending on the tab it was listened from? Are there statistically significant differences?

We saw from the previous analysis that ‘My library’ has a higher replay probability that the rest of the tabs. We wanted to analyzed all tabs individually to see if we can spot any other interesting differences.

For that purpose, we computed the probability of a song to be listened again depending on the tab the song was listened from. We calculated 95% confidence intervals of the probabilities. Here are the results:

- Explore tab: [0.420, 0.425]
- My library tab: [0.619, 0.620]
- Search tab: [0.420, 0.423]
- Discover tab: [0.415, 0.416]
- Radio tab: [0.221, 0.224]
- Listen with tab: [0.325, 0.329]
- Notification tab: [0.366, 0.390]
- Setting tab: [0.570, 0.611]

Confidence intervals were computed using the frequentist statistical approach. The following steps were performed for each tab sample:

- Compute Bernoulli probability
- Check data is large enough to be able to apply z-scores
- Compute standard error
- Compute z-score for two tailed 95% confidence level
- Compute margin of error
- Compute confidence interval based on probability and margin of error

From this we can conclude that there are clear differences in probability depending where the song was launched from, the best tabs are ‘My library’ and ‘Settings’ and the worse one is ‘Radio’.

The lower limits of the confidence interval for 'my library' tab (0.619) and for 'settings' tab (0.570) are higher that the upper limit of the rest of the tabs. The higher limit of the confidence interval of ‘radio’ tab is lower than any of the lower limits of the rest of the tabs confidence intervals.

Question 3: Are songs which belong to less genres replay more than songs which belong to more genres?

In the initial exploratory analysis we tried to check if there was a relationship between the number of genres a song belongs to and the probability of it to be played again. It looks like the more genres a song belongs to, the less probability will have to be replay. Something that could be a bit counterintuitive. We did a hypothesis test to confirm this result.

Our null hypothesis: there is no correlation between the number of genres a song belongs to and the replay probability.

H0: $\rho=0$ Ha: $\rho\neq0$

We used as test statistic the Pearson correlation ρ between both variables.

In order to perform our hypothesis test we did the following steps:

- Compute the average probability of each unique song to be replayed
- Compute the observed Pearson correlation between the replay probability of the songs and the number of genres they belong to
- Create permutation replicates assuming there is no correlation between the variables. The variable replay probability was not modified and the variable genre count was permuted.
- Compute p-value as the % of times the replicates were at least as extreme as the observed correlation.

We obtained a p-value of 1. Since the p_value is really high, we reject the null hypothesis in favor of the alternative one. There is a negative correlation between the number of genres a song belongs to and the probability of that song to be replayed.

Nonetheless the observed correlation (-0.021) is very small and therefore is not useful for practical purposes.

Question 4: Is there a significant difference between males and females in the probability of replaying songs?

We did a hypothesis test to check if there is significant difference between males and females.

The null hypothesis: the probability of songs to be replayed is the same for women and men.

H0: $\mu w=\mu m$ Ha: $\mu w\neq\mu m$

We used as test statistic the difference in means.

In order to perform our hypothesis test we did the following steps:

- Compute the average probability of each unique user to replay songs taking into account all the songs a user has listened to.
- Compute the difference of mean probability of replaying songs between the male and female sample.
- Generate permutation replicates of the difference in means. In order to generate the replicates, female and male samples are concatenated, then they are permuted and divided into two new samples.
- Compute p-value as the % of times the permutation replicates are at least as extreme as the observed difference of means.

The observed probability in men is 48.94% and in women is 47.66%.

We obtained a p-value of 0.9998. At a confidence level of 95%, provided that this is a two tailed test, the null hypothesis would be rejected. The p-value is higher than 97.5%. The observed difference in means is -1.2% between females and males. We conclude that the difference is statistically significant and men have a higher probability of replaying songs.

Question 5: Do younger users replay songs more often than older users?

We did a hypothesis test to check if user's age and the probability of listening a song again are correlated variables.

Null hypothesis: there is no correlation between the user's age and the replay probability.

H0: $\rho=0$ Ha: $\rho\neq0$

The test statistic is the Pearson correlation between both variables.

In order to perform our hypothesis test we did the following steps:

- Compute the observed Pearson correlation between the mean replay probability of songs per user and the user's age
- Create permutation replicates assuming there is no correlation between the variables. The variable replay probability was not modified and the variable age was permuted.
- Compute p-value as the % of times the replicates were at least as extreme as the observed correlation.

The observed Pearson correlation between the variables is -0.09.

We obtained a p-value of 1. Since the p_value is really high, we reject the null hypothesis in favor of the alternative one.

There is a week negative correlation between the users' age and the mean probability of a user to replay songs. Overall the probability or replaying songs decreases as users get older.