Data Analysis Report

Haohang Yan

Before doing the analysis, we remove NaN data element wise and put the cleaned ratings of movie into matrix M1.

Question 1

We split the data by the median of the popularity (rating count). Rating count below median of popularity is considered less popular and rating count above median of popularity is considered popular. Then we add ratings of movies into two splitted group.

Null Hypothesis: Movies that are more popular are not rated higher than movies that are less popular

Alternative Hypothesis: Movies that are more popular are rated higher than movies that are less popular

We use a one tail Mann-Whitney U test. Mann-Whitney is nonparametric. Two samples are independent, and their distribution is not normally distributed. It is skewed because in our data we have 807 female, 260 male and 6 self-described genders. It does not represent the whole population. (In Question 1-9, the reasons of using Mann-Whitney U test are the same)We choose one-tailed because we want to know if sample1 is distributed left to sample2

After applying Mann-Whitney test, the p value is 0, less than 0.005. we reject the null hypothesis and conclude that **movies that are more popular are rated higher than movies that are less popular.**

Question 2

We split the data by the median of the year the movie was release and then add the ratings of each movie to list old_movies_rating and new_movies_rating.

Null Hypothesis: Movies that are newer are not rated differently than movies that are older

Alternative Hypothesis: Movies that are newer are rated differently than movies that are older

After applying Mann-Whitney U test, the p value is 0.2269 which is greater than 0.005. The null hypothesis is true and we conclude that **movies that are newer are rated differently than movies that are older**

Question3

We split the rating of Shrek(2001) by gender and generate two list.

Null Hypothesis: Male and female viewers rate it the same

Alternative Hypothesis: Male and female views rate it differently

After applying two-tailed Mann-Whitney U test, the p value is 0.0505 which is greater than 0.005. The null hypothesis is true and we conclude that **Male and female viewers rate it the same**

Question 4

We iterate the data by movie names and perform two-tailed Mann-Whitney U test. Same step in question3. Then we count the number of tests which has p value less than 0.005(means null hypotheses is rejected and male and female viewers rate it differently) and then divided by total number of movies. The result is **0.125(12.5%)**

Question5

We split the rating of The Lion King (1994) by "Only Child" and generate two list. One is rating by people who are the only child and another one is rating by people who have siblings.

Null Hypothesis: People who are only children do not enjoy 'The Lion King (1994)' more than people with siblings

Alternative Hypothesis: People who are only children enjoy 'The Lion King (1994)' more than people with siblings. After applying one-tailed Mann-Whitney U test, the p value is 0.9784 which is greater than 0.005. We cannot reject the Null hypothesis and conclude that **People who are only children do not enjoy 'The Lion King (1994)' more than people with siblings**

Question 6

We iterate the data by movie names and perform two-tailed Mann-Whitney U test. For each movie we make hypothesis:

Null Hypothesis: The movie is rated the same by viewers with siblings vs. those without

Alternative Hypothesis: The movie is are not rated the same by viewers with siblings vs. those without

we count the number of tests which has p value less than 0.005(means null hypotheses is rejected and The movie is rated different by viewers with siblings vs. those without) and then divided by total number of movies. The result is **0.0175(1.75%)**

Question 7

We split the rating data from The Wolf of Wall Street (2013) by if the person enjoy watching the movie alone. One is rating by people enjoy watch movie alone and another one is rating by people who do not like to watch movie alone.

Null Hypothesis: people who like to watch movies socially do not enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone

Alternative Hypothesis: people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone

After applying one-tailed Mann-Whitney U test, the p value is 0.9437 which is greater than 0.005. We cannot reject the Null hypothesis and conclude that **people who like to watch movies socially do not enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone**

Question 8

We iterate the data by movie names and perform two-tailed Mann-Whitney U test. Then we count the number of tests which has p value less than 0.005(means null hypotheses is rejected and the movie exhibit a "social watching" effect) and then divided by total number of movies. The result **is 0.015(1.5%).**

Question 9

We take data from movie 'Home Alone (1990)' and 'Finding Nemo (2003)' into two list and remove nan values. And then perform two-tailed Mann-Whitney U test

Null Hypothesis: Ratings distribution of 'Home Alone (1990)' are the same as that of 'Finding Nemo (2003)'

Alternative Hypothesis: Ratings distribution of 'Home Alone (1990)' are different from that of 'Finding Nemo (2003)'

After applying the test, we found that the p value is far smaller than 0.005. We reject the null hypothesis and conclude that **ratings distribution of 'Home Alone (1990)' are different from that of 'Finding Nemo (2003)'**

Question 10

We iterate over franchise name and find movies in this franchise and put the data in f_data. I remove NaN element wise for each movie and then perform Kruskal-Wallis H-tes. This test works on 2 or more independent data and we know that the data population are not normally distributed with different sizes.(for example, the data is skewed by gender, also mentioned in question 1). Then we make hypothesis.

Null Hypothesis: Movie quality are consistent

Alternative Hypothesis: Movie quality are not consistent.

After applying Kruskal-Wallis test, we get the p value, if the p value is less than 0.005, we reject the null hypothesis and conclude that movie quality are not consistent.

Finally, we conclude that only “Harry Potter” franchise has consistent quality, while others are not. There are **7 of them are of inconsistent quality**.

Extra

We want to know the percentage of the people who are most likely to cry (I have cried during the movie=5) rate movies differently than people are less likely to cry(I have cried during the movie=1). We set the significant level to 0.005 and perform Mann Whitney U test.

We get the result that 2% of the movies are rated differently by people who are most likely to cry and people who are less likely to cry.