Data Analysis Project 1

Hypothesis testing of movie ratings data

Luigi Noto

– Question 1

In order to assess whether more popular movies are rated higher than less popular movies, I computed the popularity of each movie, which is defined as the number of ratings the movie has received, and computed the median of the popularity measures of each movie, which turned out to be equal to 197.5. Then, I collected in a list the titles of all the movies with popularity greater than or equal to 197.5, considered as the "more popular" movies, and in another list the titles of all the other movies, considered as the "less popular" ones. I indexed the ratings dataframe with each of these lists, ending up with a more popular movies dataframe and a less popular movies dataframe. Since the question asked for a comparison in ratings between these two categories of movies, I combined all the ratings for the popular movies in an array and all the ratings for the less popular movies in another array, ending up with two samples of ratings. After removing the missing data in each of the two samples, i.e. element-wise, in order to compare the two samples and make inference about whether the two samples come from the same underlying distribution, since for rating data it is not recommended to compare sample means, I performed a one-sided Mann Whitney U test, which compares sample medians. The alternative hypothesis is that the underlying distribution F of more popular movies is stochastically greater than the underlying distribution G of less popular movies, meaning that F has higher median than G. The p-value is (about) 0.0. Thus, there is no statistical significance at $\alpha = 0.005$, meaning that I rejected the null hypothesis and concluded that movies that are more popular are rated higher than movies that are less popular.

– Question 2

In order to assess whether newer movies are rated differently than older movies, I retrieved the release year of each movie from its title (except for 'Rambo: First Blood Part II', that did not have the release year in the title, which I looked up on Wikipedia). Then, I computed the median of the release years of all the movies, which turned out to be equal to 1999, and performed a median-split of the movies based on the release year, ending up with a new movies dataframe, containing the ratings for the movies released in 1999 or later, and an old movies dataframe, containing the ratings for the movies released before 1999. Since the question asked for a comparison in ratings between these two categories of movies, I combined all the ratings for the new movies in an array and all the ratings for the old movies in another array, ending up with two samples of ratings. After removing the missing data in each of the two samples, i.e. element-wise, since for rating data it is not recommended to compare sample means, I performed a two-sided Mann Whitney U test, which compares sample medians. The alternative hypothesis is that the median of the underlying distribution of new movies is different from the median of the underlying distribution of old movies. The p-value is about $1.28 \cdot 10^{-6}$. Thus, the difference in sample medians is significant at $\alpha = 0.005$, meaning that I rejected the null hypothesis and concluded that newer movies are rated differently than older movies.

– Question 3

To answer this question, I retrieved the ratings for 'Shrek (2001)' from the ratings data, and divided them into an array containing the ratings given by male viewers and an array containing the ratings given by female viewers. I then removed the missing data element-wise, given that the samples are not paired, but independent. Since for rating data it is not recommended to compare sample means, I performed a two-sided Mann Whitney U test, which compares sample medians. The alternative hypothesis is that the median of the underlying distribution of male ratings is different from the median of the underlying distribution of female ratings. The p-value is about 0.051. Thus, the difference in sample medians is not significant at $\alpha = 0.005$, meaning that I failed to reject the null hypothesis and concluded that the enjoyment of 'Shrek (2001)' is not gendered.

– Question 4

To answer this question, I followed the same procedure used in Question 3 for each movie and computed the proportion of movies for which the difference in the sample medians of male and female ratings is significant at $\alpha = 0.005$, i.e. the associated p-value is lower than 0.005, which turned out to be equal to 0.125. Thus, 12.5% of movies are rated differently by male and female viewers.

– Question 5

To answer this question, I retrieved the ratings for 'The Lion King (1994)' from the ratings data, and divided them into an array containing the ratings given by people who are only children and an array containing the ratings given by people with siblings (ignoring -1 values, which are missing data). I then removed the missing data element-wise, given that the samples are not paired, but independent. Since for rating data it is not recommended to compare sample

means, I performed a one-sided Mann Whitney U test, which compares sample medians. The alternative hypothesis is that is that the underlying distribution F of ratings of people who are only children is stochastically greater than the underlying distribution G of people with siblings, meaning that F has higher median than G. The p-value is about 0.98. Thus, there is no statistical significance at $\alpha = 0.005$, meaning that I failed to reject the null hypothesis and concluded that that people who are only children do not enjoy 'The Lion King (1994)' more than people with siblings.

- **Question 6**

  To answer this question, I followed the same procedure used in Question 5 for each movie, but with "two-sided" alternative hypothesis, i.e. performing a two-sided test, and computed the proportion of movies for which the difference in the sample medians of ratings of people who are only children and ratings of people with siblings and is significant at $\alpha = 0.005$, i.e. the associated p-value is lower than 0.005, which turned out to be equal to 0.0175. Thus, 1.75% of movies are rated differently by people who are only children and people with siblings.

- **Question 7**

  To answer this question, I retrieved the ratings for 'The Wolf of Wall Street (2013)' from the ratings data, and divided them into an array containing the ratings given by people who like to watch movies socially and an array containing the ratings given by people who prefer watch movies alone (ignoring -1 values, which are missing data). I then removed the missing data element-wise, given that the samples are not paired, but independent. Since for rating data it is not recommended to compare sample means, I performed a one-sided Mann Whitney U test, which compares sample medians. The alternative hypothesis is that is that the underlying distribution F of ratings of people who like to watch movies socially is stochastically greater than the underlying distribution G of people who prefer watch movies alone, meaning that F has higher median than G. The p-value is about 0.94. Thus, there is no statistical significance at $\alpha = 0.005$, meaning that I failed to reject the null hypothesis and concluded that 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**

  To answer this question, I followed the same procedure used in Question 7 for each movie and computed the proportion of movies for which the sample median of ratings from people who like to watch movies socially is significantly higher than the median of ratings from people who prefer to watch movies alone at $\alpha = 0.005$, i.e. the associated p-value is lower than 0.005, which turned out to be equal to 0.015. Thus, 1.5% of movies exhibit such a "social watching" effect.

- **Question 9**

  In order to assess whether the ratings distribution of 'Home Alone (1990)' is different than that of 'Finding Nemo (2003)', I collected the ratings for 'Home Alone (1990)' in an array and the ratings for 'Finding Nemo (2003)' in another array. Since we have a "within-subjects" design, i.e. repeated measures for each person, it is better to perform a test for paired samples. I then removed missing data row-wise, in order to obtain samples of the same size and to account for viewers' choices. Since the two samples are dependent, it is better not to use the Mann Whitney U test, which assumes independent samples. Instead, I performed a two-sided Wilcoxon signed-rank test, which is a non-parametric test for paired samples. It tests the null hypothesis that the median of the differences in ratings of each viewer is zero against the alternative that it is different from zero. The p-value is about $3.9 \cdot 10^{-17}$. Thus, the median of the differences in ratings is significantly different from zero, meaning that I rejected the null hypothesis and concluded that the ratings distribution of 'Home Alone (1990)' is different than that of 'Finding Nemo (2003)'.

- **Question 10**

  In order to assess how many movies of the given franchises are of inconsistent quality, I tested whether or not the ratings for all the movies of each franchise come from the same underlying distribution. For each franchise, I performed a test with the ratings of all the movies of the franchise. Since, in each of the tests, we have a "within-subjects" design, i.e. repeated measures for each person, it is better to perform a test for dependent samples. Then, for each list, I removed missing data in the arrays in the list row-wise, in order to obtain samples of the same size and to account for viewers' choices. Since the samples are dependent, it is better not to use the Kruskal-Wallis test, which assumes independent samples. Instead, for each franchise, I performed a Friedman Chi Square test with the ratings arrays of the movies in that franchise, which tests the null hypothesis that repeated measurements of the same individuals have the same distribution. The franchises of inconsistent quality are those for which the test yielded a p-value lower than 0.005. It turned out that the results of all the tests are statistically significant, meaning that I rejected the null hypothesis for all franchises and concluded that all franchises are of inconsistent quality.
  By performing element-wise removal of missing data for the ratings of the movies in each franchise and using the Kruskal-Wallis test, which assumes independent samples, we would conclude that all the franchises except for 'Harry Potter' are of inconsistent quality.