DATA ANALYSIS PROJECT 1

Q.1). Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?

D (Do): I did a median split on the number of ratings provided for each movie and judged the movie's popularity by comparing to this median split value. I computed the median rating of each movie and stated my null hypothesis: "Popular movies are not rated higher than less popular ones." I then did a Mann Whitney U significance test.

Y (Why/Reasoning): I took the median of ratings as it is not reasonable to take the mean of ratings since there are psychological, not purely mathematical, differences between user ratings. The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal.

F (Find): I obtained a test statistic value of U = 33427.5 from the Mann Whitney U significance test. This gave me a p-value of 1.986e-34, which means the probability of the null hypothesis assumption being true is 1.986e-34.

A (Answer): Since the p-value is less than α = 0.005, this value is significant and we reject the assumption that popular movies are not rated higher than less popular ones.

Q.2). Are movies that are newer rated differently than movies that are older?

D (Do): I did a median split on the release years provided for each movie and judged the movie's "age" by comparing to this median split value. I computed the median rating of each movie and stated my null hypothesis: "Newer movies are not rated higher than older ones." I then did a Mann Whitney U significance test.

Y (Why/Reasoning): I took the median of ratings as it is not reasonable to take the mean of ratings since there are psychological, not purely mathematical, differences between user ratings. The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal.

F (Find): I obtained a test statistic value of U = 18127.5 from the Mann Whitney U significance test. This gave me a p-value of 0.0887, which means the probability of the null hypothesis assumption being true is 0.0887.

A (Answer): Since the p-value is greater than α = 0.005, this value is not significant and we cannot reject the assumption that newer movies are not rated higher than older ones.

Q.3).Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?

D (Do): I filtered the data to only observe the ratings of Shrek, and further filtered it by gender to only include female(1) and male(2) genders. I also did an element-wise drop of the null values. I stated my null hypothesis: “Males and females do not rate Shrek differently.” and then did a Mann Whitney U significance test.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal. I handled the null values so as to not lose out on power.

F (Find): I obtained a test statistic value of U = 82232.5 from the Mann Whitney U significance test. This gave me a p-value of 0.0505, which means the probability of the null hypothesis assumption being true is 0.0887.

A (Answer): Since the p-value is greater than α = 0.005, this value is not significant and we cannot reject the assumption that males and females do not rate Shrek differently.

Q.4). What proportion of movies are rated differently by male and female viewers?

D (Do): I gathered all the ratings given by men and women for each movie and obtained p-values for each movie by running them through a Mann Whitney U Test. My null hypothesis was “Movies are not rated differently by males and females” and so I counted those values that rejected this null hypothesis assumption and divided the count by the total number of p-values to obtain the desired proportion.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal.

F (Find): I obtained a proportion of 0.125.

A (Answer): 12.5% of the movies are rated differently by males and females.

Q.5). Do people who are only children enjoy ‘The Lion King (1994)’ more than people with siblings?

D (Do): I filtered the data to only observe the ratings of The Lion King, and further filtered it by the column ‘Are you an only child?...’ to exclude those who did not respond. I also did an element-wise drop of the null values. I stated my null hypothesis: “Only children do not enjoy The Lion King ” more than people with siblings.” and then did a Mann Whitney U significance test.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal. I handled null values so as to not lose out on power.

F (Find): I obtained a test statistic value of U = 52929.0 from the Mann Whitney U significance test. This gave me a p-value of 0.0432, which means the probability of the null hypothesis assumption being true is 0.0432.

A (Answer): Since the p-value is less than α = 0.005, this value is not significant and we cannot reject the assumption that only children do not enjoy The Lion King more than people with siblings.

Q.6). What proportion of movies exhibit an “only child effect”, i.e. are rated different by viewers with siblings vs. those without?

D (Do): I gathered all the ratings given by only children and viewers with siblings for each movie and obtained p-values for each movie by running them through a Mann Whitney U Test. I also did an element-wise drop of the null values. My null hypothesis was “Movies are not rated differently by only children and viewers with siblings” and so I counted those values that rejected this null hypothesis assumption and divided the count by the total number of p-values to obtain the desired proportion.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal. I handled null values so as to not lose out on power.

F (Find): I obtained a proportion of 0.0175.

A (Answer): 1.75% of the movies are rated differently by only children vs. viewers with siblings.

Q.7). Do people who like to watch movies socially enjoy ‘The Wolf of Wall Street (2013)’ more than those who prefer to watch them alone?

D (Do): I filtered the data to only observe the ratings of ‘The Wolf of Wall Street’, and further filtered it by the column ‘Movies are best enjoyed alone…’ to exclude those who did not respond. I also did an element-wise drop of the null values. I stated my null hypothesis: “People who like to watch movies socially do not enjoy ‘The Wolf of Wall Street’ more than those who prefer to watch them alone” and then did a Mann Whitney U significance test.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal. I handled null values so as to not lose out on power.

F (Find): I obtained a test statistic value of U = 56806.5 from the Mann Whitney U significance test. This gave me a p-value of 0.1128, which means the probability of the null hypothesis assumption being true is 0.1128.

A (Answer): Since the p-value is greater than α = 0.05, this value is not significant and we cannot reject the assumption that people who like to watch movies socially do not enjoy ‘The Wolf of Wall Street’ more than those who prefer to watch them alone.

Q.8). What proportion of movies exhibit such a “social watching” effect?

D (Do): I gathered all the ratings given by people who watch movies socially and those who prefer to watch them alone, for each movie and obtained p-values for each movie by running them through a Mann Whitney U Test. I also did an element-wise drop of the null values. My null hypothesis was “Movies are not rated differently by people who watch movies socially and those who watch them alone” and so I counted those values that rejected this null hypothesis assumption and divided the count by the total number of p-values to obtain the desired proportion.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal. I handled null values so as to not lose out on power.

F (Find): I obtained a proportion of 0.025.

A (Answer): 2.5% of the movies are rated differently by people who like to watch movies socially vs. viewers who like to watch movies alone.

Q.9). Is the ratings distribution of 'Home Alone (1990)' different from that of 'Finding Nemo (2003)'?

D (Do): I obtained all the ratings of 'Home Alone' and 'Finding Nemo' and performed a Kruskal-Wallis Test on both sets of ratings. I also did an element-wise drop of the null values. My null hypothesis was "The ratings distribution of 'Home Alone' is not different from that of 'Finding Nemo'" .

Y (Why/Reasoning): The reasoning behind conducting the Kruskal-Wallis test is because we are not dealing with a comparison of means or medians, rather we want to test if the underlying distributions of two samples is the same. I handled null values so as to not lose out on power.

F (Find): I obtained a test statistic of 0.1431 and a corresponding p-value of 3.2626e-10. This means that the probability of the null hypothesis assumption being true is 3.2626e-10.

A (Answer): Since the p-value is less than α = 0.005, this value is significant and we reject the null hypothesis assumption that the ratings distribution of 'Home Alone' is not different from that of 'Finding Nemo'.

Q.10).There are ratings on movies from several franchises (['Star Wars', 'Harry Potter', 'The Matrix', 'Indiana Jones', 'Jurassic Park', 'Pirates of the Caribbean', 'Toy Story', 'Batman']) in this dataset. How many of these are of inconsistent quality, as experienced by viewers?

D (Do): I gathered the ratings for each movie belonging to each franchise. I, then, ran each set of ratings through the Kruskal Wallis significance test and obtained p-values for each franchise. I also did an element-wise drop of the null values. My null hypothesis was "The movies in the franchise have the same underlying distribution" and so I observed which franchise did not have the same underlying distribution and consequently had inconsistent quality as expressed by its viewers.

Y (Why/Reasoning): To judge inconsistent quality as expressed by viewers, we can compare the underlying distributions of all the movie ratings of each franchise. The reasoning behind conducting the Kruskal-Wallis test is because we are not dealing with a comparison of means or medians, rather we want to test if the underlying distributions of two samples is the same, and also because we are dealing with more than 3 groups. I handled null values so as to not lose out on power.

F (Find): I obtained test statistics of 230.584, 3.3312, 48.3789, 45.7942, 46.5909, 20.6440, 24.3860, and 190.5350 with corresponding p-values of 8.016e-48, 0.3433, 3.1237e-11, 6.2728e-10, 7.6370e-11, 3.2901e-05, 5.0659e-06, and 4.2253e-42 for the franchises of 'Star Wars', 'Harry Potter', 'The Matrix', 'Indiana Jones', 'Jurassic Park', 'Pirates of the Caribbean', 'Toy Story', and 'Batman' respectively.

A (Answer): From the obtained p-values we can see that the p-value of ‘Harry Potter’ is the only one that is greater than α = 0.005, which means it is not significant and we cannot reject the null hypothesis for this franchise that the movies in this franchise have the same underlying distribution, therefore it is of an inconsistent quality.

– Extra Credit

D (Do): I filtered the data to observe the ratings for The Conjuring and people who are emotionally stable and don’t get upset easily. Since this is a well-known and well-liked horror movie, I expected stable people to enjoy it. I also did an element-wise drop of the null values. I stated my null hypothesis: “People who are emotionally stable did not rate The Conjuring highly.” and then did a Mann Whitney U significance test.

Y (Why/Reasoning): The reasoning behind conducting the Mann Whitney U test is because we are dealing with a variable that is not distributed normally and is ordinal. I handled null values so as to not lose out on power.

F (Find): I obtained a test statistic value of U = 11328.0 from the Mann Whitney U significance test. This gave me a p-value of 2.0621e-19, which means the probability of the null hypothesis assumption being true is 2.0621e-19.

A (Answer): Since the p-value is less than α = 0.005, this value is significant and we can reject the assumption that emotionally stable people did not rate ‘The Conjuring’ highly.

**** I have attached the code file which contains all the different plots for each question of this project.**