Allison Redfern
DS-GA1001 Data Analysis Project 1: Hypothesis testing of movie ratings data

Reasoning and Assumptions Behind Mann-Whitney U Test (Questions 1- 8 & Bonus): The following analyses were conducted using the Mann-Whitney U test. This test was chosen over its parametric counterparts (Independent Samples or Welch's t test) because when dealing with movie ratings, the fact that they are ranked values does not allow reduction to the sample mean for hypothesis testing. The psychological difference between two rating values is not necessarily even, and therefore shouldn't be tested as such. The null hypothesis of the Mann-Whitney U test is that the two samples come from the sample population with the same median. A significant result from this test in terms of movie ratings means that the two movies compared were rated differently. By using this test we assume that the ratings are independent from each other and that the distributions between each group are the same shape.

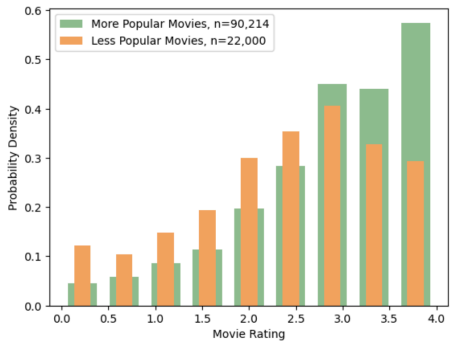

Figure 1: Movie Ratings by Popularity

1: Are movies that are more popular rated higher than movies that are less popular? To test whether popular movies are rated higher, a Mann-Whitney U test was used to see if ratings from popular and not popular movies come from a population with the same median. Popular vs not popular movies were split by the median count of ratings for each movie (greater than or less than 197.5), and an element-wise removal was performed on missing data, since the responses between the two groups are not linked. The Mann-Whitney U test resulted in test statistics, $U_{Popular}$ = 1242808144.5 & $U_{Not\ Popular}$ = 741899855.5, and a p-value so low that it only can be shown as 0.0. Therefore, we can conclude that more popular movies are rated differently (and by Figure 1, we can tell "higher") than movies that are less popular. This was a very high-powered test due to the large sample size, and therefore there are no concerns when making this inference.

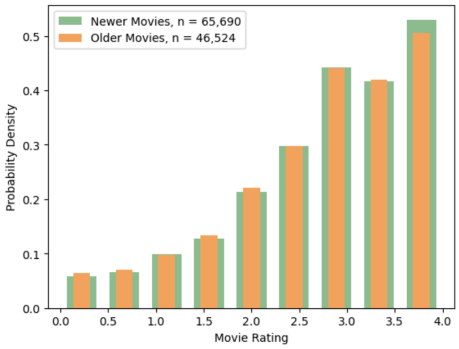

Figure 2: Movie Ratings by Age

2: Are movies that are newer rated differently than movies that are older? A Mann-Whitney U test was used to see if newer movies come from a population with the same median as older movies. A median split was performed across all movie ratings to categorize them as newer or older. The median year was 1999. Since there are multiple movies in the data set from 1999, all 1999 movies were included in the newer movie group to provide a more even split across the 400 movies. 203 movies were categorized as "newer" and 197 movies were categorized as "older". An element-wise removal was performed on missing data. The Mann-Whitney U test resulted in a test statistics, U_{Newer} = 1553577699.0 & U_{Older} = 1502583861.0, and a p-value of 1.2849×10^{-6}, which is well below the set α-level of 0.005. Therefore, it can be inferred that newer movies are rated differently than older movies. This analysis was an interesting case of statistical power. Upon looking at Figure 2, one may assume that older and newer movies are rated the same, but because the sample size is so large, the U test has much greater power and is able to detect tinier differences.

3: Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently? A Mann-Whitney U test was used to see if male and female ratings of "Shrek" come from a population with the same median. The results of Mann-Whitney U-test comparing the female and the male data resulted in test statistics, U_{Female} = 96830.5 & U_{Male} = 82232.5, and a p-value of 0.0505. Since this is not below the α-level of 0.005, the null hypothesis cannot be rejected, therefore we cannot conclude that females and males rate "Shrek" differently. The concern for this test is a matter of sample size. Looking at the differences in Figure 3 and the fact that the medians for female and male respectively were 3.5 and 3, with a higher powered test with a greater sample size (specifically of males), we may be able to see if there are significant differences in ratings between groups.

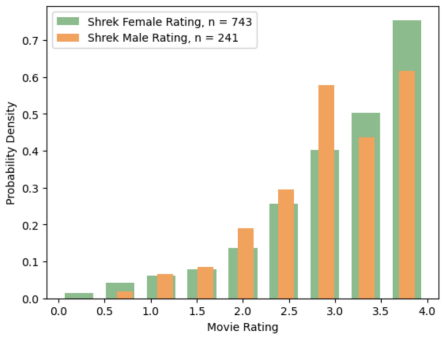

Figure 3: Shrek Ratings by Gender

4: What proportion of movies are rated differently by male and female viewers? 400 Mann-Whitney U tests were completed across all movies to see if male and female ratings for each movie come from a population with the same median. The tests resulting in p-values less than 0.005 were categorized as significant (the movies are rated differently), and others were categorized as not rated differently. From this analysis, we conclude that 50 out of 400 or $\frac{1}{8}$ (see Figure 4) of the movies were rated differently by females and males. This result assumes that each iteration of the U test had a high enough sample size for high-powered results, which may not be the case for less popular movies, potentially leading to false significant or not significant results.

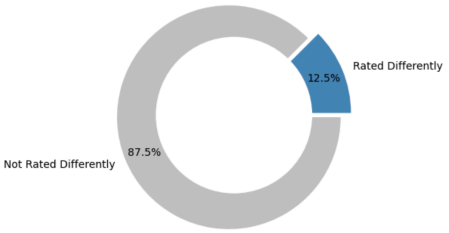

Figure 4: Proportion of Movies Rated Differently by Males & Females

5: Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings? A Mann-Whitney U test was used to see if movie ratings for "The Lion King" from people that only-children come from a population with the same median as those with siblings. The Mann-Whitney U test resulted in test statistics, $U_{Only\ Child}$ = 52929.0 & $U_{Has\ Siblings}$ = 64247.0, and a p-value of 0.0432, which is greater than α. It can be inferred that only-children and people with siblings do not rate "The

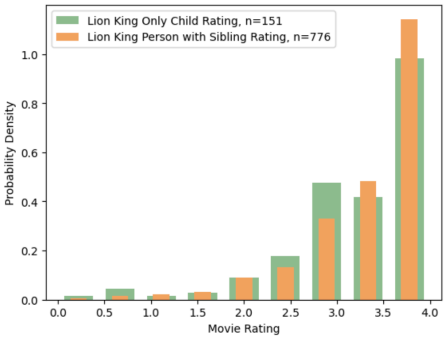

Figure 5: Lion King Ratings by Only Child or Not

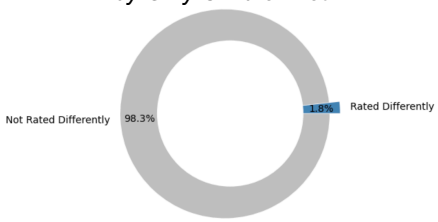

Figure 6: Proportion of Movies Rated with "Only Child Effect"

Lion King" differently. Figure 5 shows that there may be some differences. Therefore, this again was a lower powered test, so we may get more significant results if redone with a higher sample size especially with more people that are only-children.

6: What proportion of movies exhibit an "only child effect", i.e. are rated different by viewers with siblings vs. those without? 400 Mann-Whitney U tests were completed across all movies to see if only-children ratings for each movie come from a population with the same median as ratings from people with siblings. The tests that resulted in p-values less than 0.005 were then categorized as significant (the movies are rated differently), and the others were categorized as not rated differently. From this analysis, we conclude that only $\frac{7}{400}$ (see Figure 6) of the movies were rated differently by only-children vs those with siblings. This result again assumes that each U test had a high enough sample size for high-powered results, which may not be the case for all 400 tests.

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? A Mann-Whitney U test was used to see if "The Wolf of Wall Street" ratings from people who enjoy watching movies socially come from a population with the same median as those who prefer watching movies alone. The Mann-Whitney U test resulted in test statistics, U_{Alone} = 56806.5 & U_{Social} = 49303.5, and a p-value of 0.1128, which is above the set α-level. Therefore, it cannot be concluded that people who enjoy watching movies with others rate "The Wolf of Wall Street" differently than others. This once again was a lower-powered test, and with a greater sample size, we may see more significant results or the differences shown in Figure 7 may diminish.

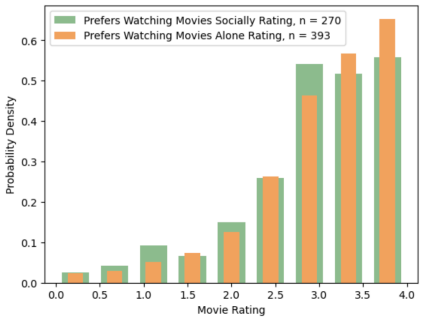

Figure 7: The Wolf of Wall Street Ratings by Alone vs Social Watcher

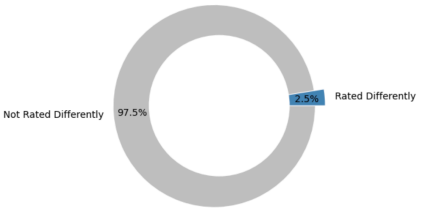

Figure 8: Proportion of Movies Rated with "Social Watching Effect"

8: What proportion of movies exhibit such a "social watching" effect? Using the same strategy as questions 4 & 6, Mann-Whitney U tests were iterated over the 400 movies to assess if ratings from people who prefer watching movies alone come from a population with the same median as people who are "social watchers". Each test that resulted in p-values less than 0.005 were then categorized as significant (the movies are rated differently), and the others were categorized as not rated differently. From this analysis, we conclude that 10 out of 400 or $\frac{1}{40}$ (see Figure 8) of the movies were rated differently by people who prefer watching movies alone vs with others. Again, this result assumes that each U test had a high enough sample size for high-powered results, which may not be the case for movies that have less ratings overall.

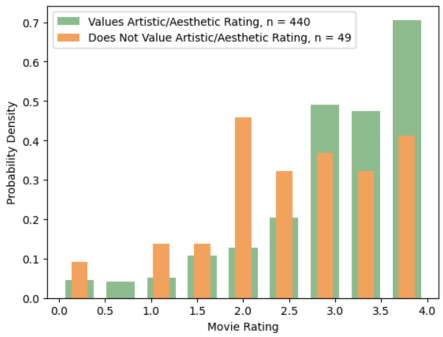

Figure 9: The Nightmare Before Christmas by Artistic Value

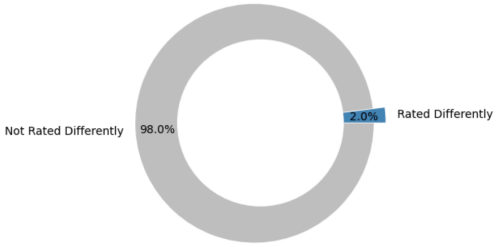

Figure 10: Proportion of Movies Rated Differently by Value of Art/Aesthetics

Bonus: Do people who value artistic/aesthetic experiences rate 'The Nightmare Before Christmas (1993)' differently than those who do not? What is the proportion of movies that are rated differently by those who value artistic/aesthetic experiences? By using a Mann-Whitney U test, we can determine if the ratings from "The Nightmare Before Christmas" from people that value artistic and aesthetic experiences come from a population with the same median as those who do not. To categorize the data we are assuming that those who responded 4 or 5 to the question of "values artistic/aesthetic experiences" value these experiences, and 1 or 2 responses do not. Neutral responses of 3 were not included in either set. The Mann-Whitney U test resulted in test statistics, $U_{Artistic}$ = 13701.0 & $U_{Not\ Artistic}$ = 7859.0, and a p-value of 0.0014. The p-value is less α so we can conclude that people who value artistic and aesthetic experiences rate "The Nightmare Before Christmas" differently than those who do not. However, in Figure 9, we can see that the sample size of those who do not value artistic/aesthetic experiences is very small, so to be confident and have a higher-powered result, it would be necessary to repeat this test with a higher sample size.

Similar to questions 4, 6, and 8 we can iterate 400 Mann-Whitney U tests across all movies to see if ratings from people who value artistic/aesthetic experiences come from a population with the same median as people who do not. Those tests with a p-value less than 0.005 were considered to be rated differently. Only 8 out of 400 or $\frac{1}{50}$ movies were enjoyed differently (shown in Figure 10). This is an interesting result because 98% of the movies tested are enjoyed by a wider audience equally rather than only being able to be enjoyed (or equally not enjoyed) by people with higher artistic and aesthetic standards.

Reasoning and Assumptions Behind Kolmogorov-Smirnov (KS) Test (Question 9): The Kolmogorov-Smirnov (KS) test was chosen for the following analysis since we are comparing distributions of ratings. The KS test compares distributions of two samples by comparing the largest point of separation between each sample's cumulative distribution functions (CDF). The null hypothesis for the KS test is that both data samples come from the same population. A significant result from this test in terms of movie ratings means that the two movies come from populations with different distributions. When using the KS test, we are assuming that the ratings are all independent from each other and that since this data is ordinal and not continuous, the results will not be exact and instead will be conservative. Also, KS tests tend to be much more sensitive toward the center of the distribution than the tails of the distribution.

9: Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'? With the assumptions stated above, the KS test was used to compare the ratings distributions of "Home Alone" and "Finding Nemo" by comparing their cumulative density functions. Because we are only interested in comparing distributions and do not care if all respondents had seen both movies, missing data was removed element-wise. The KS test resulted in test statistic (or the largest separation between the CDFs), D = 0.1527, and a p-value of 6.3794×10^{-10}, well below the set α-level of 0.005. Therefore, we conclude that the rating distributions of "Home Alone" and "Finding Nemo" are different. This difference can be seen visually in the histogram in Figure 11. Since there is a substantial sample size for this test, there are not major concerns with this result.

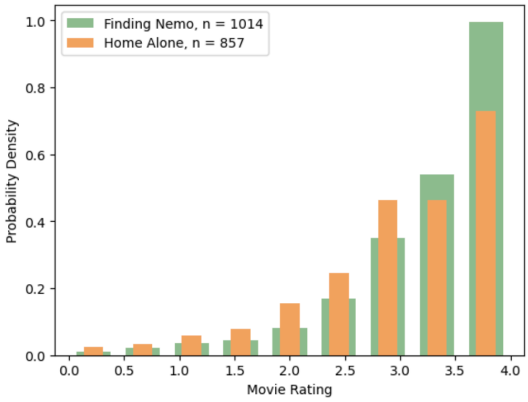

Figure 11: Finding Nemo and Home Alone Rating Distribution Comparison

Reasoning and Assumptions Behind Kruskal-Wallis Test (Question 10): The following analysis was conducted using a Kruskal-Wallis test which is an extension of the Mann-Whitney U test, used to compare medians of more than two samples. It is the non-parametric equivalent to the ANOVA test and was chosen (for the same reason the Mann-Whitney U test) because we are comparing ratings that cannot be reduced to their sample means. The null hypothesis of the Kruskal-Wallis test is that each sample comes from a population with the same median. A significant result from this test in terms of movie ratings means that the at least one movie compared was rated differently. When using the Kruskal-Wallis test we are assuming that all observations are independent of each other (this assumption will actually be broken in the following analysis, explained below) and that the distributions between each group are the same shape.

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? To compare the movies within each franchise, a Kruskal-Wallis test was conducted on the movies of each. The missing data from all movies within each franchise were removed using row-wise removal. Row-wise removal was a crucial step to ensure that only respondents that had seen every movie in a particular franchise were recorded, so consistency of quality can be best assessed. The Kruskal-Wallis test was conducted resulting in test statistics and p-values as shown in Table 1. Therefore, of all eight movies, only the "Harry Potter" and "Pirates of the Caribbean" franchises were all of consistent quality because their p-values were *greater* than the set α-level of 0.005, meaning that the movies were not rated differently among each other. See Figure 12 for a histogram example of consistent quality movies (Harry Potter) and Figure 13 showing a not consistent quality example (The Matrix). (Histograms for all can be found in the Jupyter Notebook Appendix). The main concern for this analysis is the independence assumption of the Kruskal-Wallis test. Doing row-wise removal of data contradicts this assumption since a person may rate a movie within a given franchise differently, having seen the others. For the sake of this analysis, we will hold the independence assumption and say that the rating of one movie in the franchise does not affect the rating of another movie in the franchise, however holding this assumption may affect the validity of the results.

Table 1: Results from Kruskal-Wallis Test of Movie Franchises

Franchise:	Star Wars	Harry Potter	The Matrix	Indiana Jones	Jurassic Park	Pirates of the Caribbean	Toy Story	Batman
Test Statistic (D):	193.51	5.87	40.32	54.19	49.43	6.66	23.50	84.66
p-value:	6.9402×10^{-40}	0.1179	1.7537×10^{-9}	1.0201×10^{-11}	1.8492×10^{-11}	0.0358	7.9022×10^{-6}	4.1380×10^{-19}

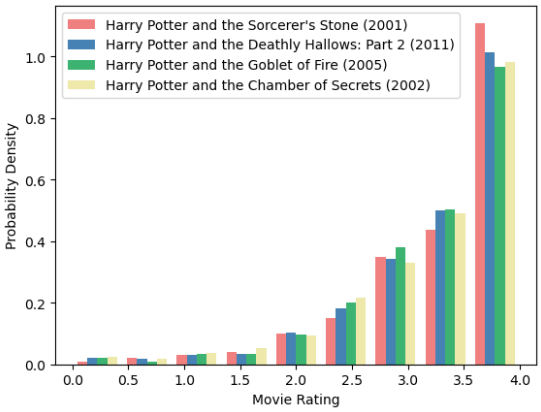

Figure 12: Harry Potter Consistent Movie Quality Ratings

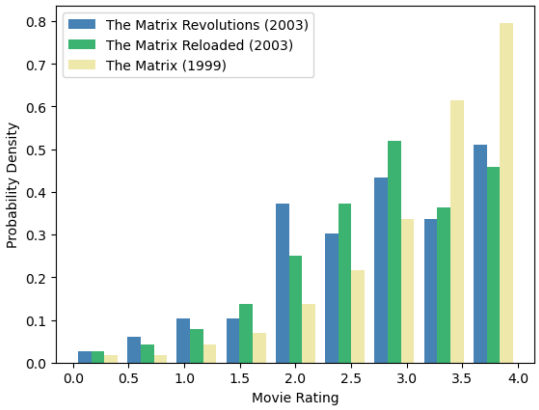

Figure 13: The Matrix Inconsistent Movie Quality Ratings