

In the movie ratings dataset, I was tasked with answering 10 questions for corporate. For each question, I performed an appropriate hypothesis test and compared the resulting p-value with our chosen significance level of 0.005. For each test, if the p-value is less than the significance level, we choose to reject the null hypothesis – that is to say, we conclude that the results are not consistent with chance alone.

Also, it does not make sense to reduce movie ratings data to their sample means, since the psychological distance between 1 star and 2 stars is not the same as the distance between 2 stars and 3 stars, for example. However, because movie ratings data are ordinal, it still makes sense to use nonparametric tests such as the Mann-Whitney U-test, since the sample medians for movie ratings still have meaning.

For questions 1-8, I use the Mann-Whitney U-test since the 2 samples for each question are independent.

- 1) A one-tailed Mann-Whitney U test produces a p-value of 9.929×10^{-35} . Therefore we reject the null hypothesis and conclude that movies that are popular are rated higher than movies that are less popular.

To obtain the 2 samples, I calculated the median total number of ratings over all movies, which is 197.5 ratings. Then each sample consisted of 200 median ratings, in which one sample had the median ratings for each movie in which its number of ratings is greater than 197.5 and the other sample had the median ratings for each movie in which its number of ratings is less than 197.5. I then performed a one-tailed U test since we are just testing whether more popular movies are rated higher than less popular movies.

- 2) A two-tailed Mann-Whitney U test produced a p-value of 0.122. Therefore we fail to reject the null, meaning we cannot conclude whether or not newer movies are rated differently than older movies.

Like the process in 1), I calculated the median date over all movies, which was 1999. In this case, I only want to compare newer movies to older movies, so I excluded in the test any movie that had a date of 1999. Then each sample consisted of median ratings in which the movies had median dates less than 1999 or greater than 1999. I then performed a two-tailed U test to see if there is any difference in the two samples.

- 3) A two-tailed Mann Whitney U test produced a p-value of 0.051. Therefore we fail to reject the null, meaning we cannot conclude whether enjoyment of Shrek (2001) is gendered.

One sample contained all the ratings of Shrek by female viewers and the other sample contained all the ratings of Shrek by male viewers.

- 4) Using a two-tailed Mann Whitney U test for all 400 movies, 12.5% of movies are rated differently by male and female viewers.

The same process in 3) was used for all 400 movies to calculate the proportion.

- 5) A one-tailed Mann-Whitney U test produced a p-value of 0.978. Therefore we fail to reject the null, meaning we cannot conclude whether only children enjoy The Lion King (1994) more than people with siblings.

One sample contained all the ratings of The Lion King by only children and the other sample contained all the ratings of The Lion King by people with siblings.

- 6) Using a two-tailed Mann-Whitney U test for all 400 movies, 1.75% of movies are rated differently by viewers with siblings and viewers without siblings.

The same process in 5) was used for all 400 movies to calculate the proportion, except that the question was just asking for any difference, so a two-tailed test was used instead.

- 7) A one-tailed Mann-Whitney U test produced a p-value of 0.943. Therefore we fail to reject the null hypothesis, meaning we cannot conclude whether The Wolf of Wall Street (2013) has a social watching effect.
- 8) Using a one-tailed Mann-Whitney U test for all 400 movies, 1.5% of movies exhibit a social watching effect.

For question 9, I used the Wilcoxon signed-rank test because many of the viewers rated both Home Alone and Finding Nemo, so the samples are dependent. Therefore a Kolmogorov-Smirnov 2-sample test would not be appropriate since a K-S 2-sample test assumes that the samples are independent. I also used row-wise removal of NaNs since the Wilcoxon signed-rank test requires the samples to have the same n.

- 9) A two-tailed Wilcoxon signed-rank test produced a p-value of 3.909×10^{-17} . Therefore we reject the null hypothesis and conclude that the ratings distributions of Home Alone (1990) and Finding Nemo (2003) are different.

After removing viewers who only watched one of the movies, one sample included all ratings of Home Alone and the other sample included all the ratings of Finding Nemo for each viewer.

For question 10, I used the Friedman test because for each franchise, many viewers rated multiple movies in the franchise. Since the Kruskal-Wallis test assumes the samples are independent, the Friedman test is much more appropriate since the Friedman test assumes that the samples are dependent, which is the case here. I also performed row-wise removal of NaNs to have the same n, and I also double checked that for each franchise, the number of data points was greater than 20. The minimum number of data points was 219 for Batman, so we are not worried about too little data.

- 10) A Friedman test for all 8 franchises resulted in 8 p-values that were all less than 0.005 (3.671×10^{-57} , 1.179×10^{-3} , 2.072×10^{-15} , 5.543×10^{-18} , 3.116×10^{-17} , 1.773×10^{-5} , 5.839×10^{-13} , 5.583×10^{-23} respectively), so we reject the null hypothesis for all 8 franchises and conclude that all 8 franchises are of inconsistent quality.

In addition to the above 10 questions, I wanted to find out whether Column 471, do emotions rub off on the viewer, impacted movie ratings, as personally I would rate myself a 6 for this question. For all 400 movies, I divided the ratings for each movie into 2 samples, one with ratings where the movie experience rating was equal to 4, 5, or 6 (higher) and one with ratings where the movie experience rating was equal to 1, 2, or 3 (lower). I found a high proportion of significant p-values (using a one-tailed Mann-Whitney U-test) of 31%, which means that for 124 out of the 400 movies, we reject the null hypothesis and conclude that people who are more emotional during movies tend to enjoy the movie more than people who aren't.

I also wanted to see if there was such an effect on Life is Beautiful (1997), one of the most emotional movies I have ever watched. After performing a one-tailed Mann-Whitney U-test, I got a p-value of 4.277×10^{-3} , which means a viewer rating highly on emotional "rubbing" increases the viewer's rating of Life is Beautiful. This supports the idea that movies that elicit strong emotional responses like Life is Beautiful are rated higher by people who are more moved by movies.