Data analysis project 1:

Hypothesis testing of movie ratings data

Submitted by Doma Ghale

1) To check if movies that are more popular rated higher than movies that are less popular I performed a one-sided Mann-Whitney U rank test, which tests median between two distributions. The resulted p-value was 9.93e-35, so I rejected the null hypothesis and concluded that the movies that are more popular are rated higher than movies that are less popular. If we simply look at the histogram of two groups we can clearly see that the distributions are different. More popular movies have more high ratings compared to less popular movies.[1]

2) I performed a two-sided MannWhitney U test to check if movies that are newer rated differently than movies that are older. The p-value I got was 0.1134 so using significance test level of 0.005, we don't reject the null hypothesis and conclude that there is not a strong evidence that the movies that are newer are rated differently than movies that are older. If we look at the histogram, we can see a very slight difference in their distributions.

3)To find if enjoyment of 'Shrek (2001)' is gendered, I performed a two-sided MannWhitney U test and got a p-value of 0.050, so I concluded that there is not strong evidence that enjoyment of 'Shrek (2001)' gendered. When I plotted the histogram, I noticed a big difference in the proportion of rating of 3.

4) To observe the proportion of movies rated differently by male and female viewers, we need to extend the previous question for the movie "Shrek" to all of the other 399 movies. For each movie I performed a two-sided MannWhitney U test and grouped the movies into one if the p-value is less than 0.005 and found the ratio to be 0.125.

6) To compute the proportion of movies that exhibit an "only child effect", I performed a two-sided Mann Whitney U test for each movie and grouped the ones with significant p-values. There were only 7 out of 400 movies (0.0175) that exhibited an "only child effect". I think an extension of this problem could be to find what is so different about these movies. I thought it is particularly interesting that out of .... Star Wars episodes "The Return of the Jedi" was the only one with the effect. The effect we are seeing could still happen just by chance, so there might not be even any meaningful difference. This is a different kind of question and out of scope for this project, so I am moving on to the next question.

7) To find out if people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone, I performed a one-sided Mann Whitney U test and found a p-value of 0.94. Therefore, I conclude that there is not strong evidence that people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone.

8) To find out the proportion of movies exhibiting such a "social watching" effect, I performed the same one-sided Mann Whitney U test and grouped the movies that showed significant p-value into one and found that 6 out of 400 movies (0.015) exhibited such a "social watching" effect.

9) To check the difference in ratings distributions between 'Home Alone (1990)' vs 'Finding Nemo (2003)' I performed K-S two sample test and got a p-value of 2.20e-10. Therefore, I conclude that the ratings distribution of 'Home Alone (1990)' is different from that of 'Finding Nemo (2003)'.

---

[1] See appendix for code, graphs and detailed notes

10) To check if the movies with a franchise are inconsistent I performed Kruskal-Wallis test and found that 6 out of 8 of those franchises have inconsistent quality.

Extra Credit: For each non movie rating question, the answers to which of the questions are significantly different between gender. To account for gender being on the high end of the spectrum, for each question, I first performed a one-sided Mann Whitney U test to check if females have greater inclination, then checked if male do. Out of 77 questions, 17 questions had greater effect for females than male, and 7 questions had greater effect for male than female. Some of the behaviour expectations we put on females are that they are emotional, moody, get scared easily, and on male are that they are risk-takers and emotionally stable. To no surprise these expectations were reflected on the data. I do not believe that this data suggests anything about the biological differences in behaviours, instead I believe that it shows the attitude and internal biases or learnt differences.

The questions where female seems to be more inclined than male were:	The questions where male seems to be more inclined than female were:
- I enjoy impulse shopping - I enjoy going to large music or dance festivals - I had a sheltered upbringing - My life is very stressful - Is talkative - Is a reliable worker - Can be tense - Worries a lot - Can be moody - Is considerate and kind to almost everyone - Gets nervous easily - I have cried during a movie - When watching a movie I cheer or shout or talk or curse at the screen - When watching a movie I feel like the things on the screen are happening to me - The emotions on the screen "rub off" on me - for instance if something sad is happening I get sad or if something frightening is happening I get scared - Movies change my position on social economic or political issues - When watching movies things get so intense that I have to stop watching	- I enjoy driving fast - Have you gambled or bet for money? - Is relaxed/handles stress well - Is emotionally stable/not easily upset - is inventive - Can be cold and aloof - Remains calm in tense situations