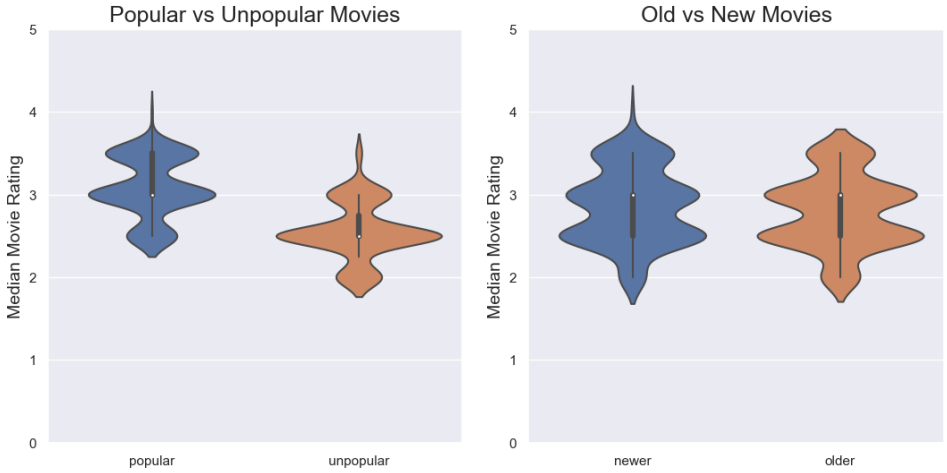
Data Analysis Project 1
Hypothesis testing of movie ratings data
Written by Mrugank Dake

The dataset is movie rating data for 400 movies. In addition, the responders have also responded to other questions pertaining to sensation-seeking behaviors, personality questions, self-reported movie experience ratings, gender identity, whether they are the only child or had siblings, and whether they tend to enjoy movies alone. Since the data is rating, it belongs to the Likert scale. Therefore, everywhere non-parametric tests have been used. Since the assignment consists of several hypotheses tests, and as suggested in the assignment instructions, a stringent alpha level of α = 0.005 is used for all hypotheses tests.

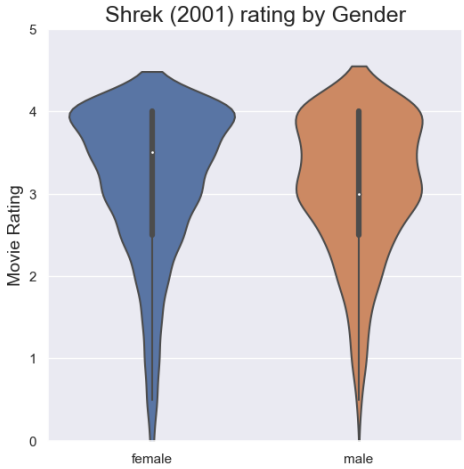
Q1. Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?
We divide the movies into popular (with higher number of ratings) (N_popular = 200) and unpopular (N_unpopular = 200) (with lower number of ratings) splitting at median number of ratings. Then we compute the median ratings for popular and unpopular movies.
H0: Movies that are more popular are not rated higher than movies that are less popular.
Ha: Movies that are more popular are rated higher than movies that are less popular.
Comparing the median ratings for popular and unpopular movies using right-tailed (since this is a right-tailed Ha) Mann-Whitney U test, we get U_popular = 33427.5, U_unpopular = 6572.5, p = 9.93e-35 < α.
Therefore, we conclude that Ha is more likely to be true, aka. movies that are more popular are rated higher than movies that are less popular.



Q2. Are movies that are newer rated differently than movies that are older?
We divide the movies into newer (N_newer = 203) and older (N_older = 197) splitting at median year of movie. Then we compute the median ratings for newer and older movies.
H0: Movies that are newer are not rated differently than movies that are older.
Ha: Movies that are newer are rated differently than movies that are older.
Comparing the median ratings for newer and older movies using two-tailed (since there is no directionality in Ha) Mann-Whitney U test, we get U_newer = 21863.5, U_older = 18127.5, p = 0.0887 > α.
Therefore, we conclude that H0 is more likely to be true, aka. movies that are newer are not rated differently than movies that are older.

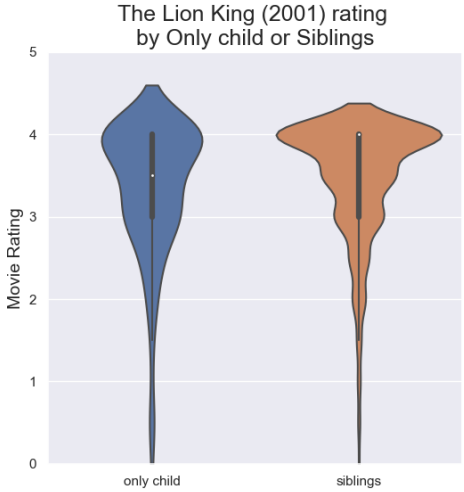
Q3. Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?
We divide the dataset into female_df (gender identity == 1) (N_female = 743) and male_df (gender_identity == 2) (N_male = 241). Then we compare the ratings for 'Shrek (2001)' for female_df and male_df.

H0: Enjoyment of Shrek (2001) is not gendered.
Ha: Enjoyment of Shrek (2001) is gendered.
Comparing the median ratings for Shrek by females and males using two-tailed (since there is no directionality in Ha) Mann-Whitney U test, we get U_female = 96830.5, U_male = 82232.5, p = 0.0505 > α.
Therefore, we conclude that H0 is more likely to be true, aka. enjoyment of Shrek (2001) is not gendered.



Q4. What proportion of movies are rated differently by male and female viewers?
Repeating the procedure in Q3 for each movie, we can determine whether p-value obtained is less than α for each movie. We find that the proportion of movies for which p < α i.e. the proportion of movies that are rated differently by male and female viewers is 12.5%.

Q5. Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?
We divide the dataset into onlychild_df (only child == 1) (N_onlychild = 151) and siblings_df (only_child == 0) (N_siblings = 776). Then we compare the ratings for 'The Lion King (1994)' for onlychild_df and siblings_df.
H0: People who are only child do not enjoy "The Lion King (1994)" more than people with siblings.
Ha: People who are only child enjoy "The Lion King (1994)" more than people with siblings.
Comparing the median ratings for The Lion King by only child and siblings using right-tailed (since this is a right-tailed Ha) Mann-Whitney U test, we get U_onlychild = 52929.0, U_siblings = 64247.0, pval = 0.9784 > α.
Therefore, we conclude that H0 is more likely to be true, aka. people who are only child do not enjoy "The Lion King (1994)" more than people with siblings.



Q6. What proportion of movies exhibit an "only child effect", i.e. are rated different by viewers with siblings vs. those without?

Repeating the procedure in Q5 for each movie but using a two-tailed test, we can determine whether p-value obtained is less than α for each movie. We find that the proportion of movies for which $p < \alpha$ i.e. the proportion of movies that are rated differently by people who are only child and those who have siblings is 1.75%.

Q7. Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?

We divide the dataset into bestalone_df (enjoy alone == 1) (N_bestalone = 393) and notbestalone_df (enjoy alone == 0) (N_notbestalone = 270). Then we compare the ratings for 'The Wolf of Wall Street (2013)' for bestalone_df and notbestalone_df.

H0: 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.

Ha: People who like to watch movies socially enjoy "The Wolf of Wall Street (2013)" more than those who prefer to watch them alone.

Comparing the median ratings for The Lion King by only child and siblings using left-tailed (since this is a left-tailed Ha) Mann-Whitney U test, we get U_bestalone = 56806.5, U_notbestalone = 49303.5, p = 0.9437 > α.

Therefore, we conclude that H0 is more likely to be true, aka. people who like to watch movies socially enjoy "The Wolf of Wall Street (2013)" more than those who prefer to watch them alone.

Q8. What proportion of movies exhibit such a "social watching" effect?

Repeating the procedure in Q7 for each movie but using a left-tailed test, we can determine whether p-value obtained is less than α for each movie. We find that the proportion of movies for which $p < \alpha$ i.e. the proportion of movies that are rated higher by people who enjoy watching movies socially is 1.5%.

Q9. Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'?

We drop the nan entries from column of 'Home Alone (1990)' (N_homealone = 857) and 'Finding Nemo (2003)' (N_findingnemo = 1014).

H0: The ratings distribution of 'Home Alone (1990)' is same as that of 'Finding Nemo (2003)'

Ha: The ratings distribution of 'Home Alone (1990)' is different than that of 'Finding Nemo (2003)'

Since this is a rating data we use Kolmogorov-Smirnov test to compare the rating distribution for the two movies. We get KS = 0.1526, pval = 6.38e-10 < α.

Therefore, we conclude that Ha is more likely to be true, aka. the ratings distribution of 'Home Alone (1990)' is different than that of 'Finding Nemo (2003)'.

Q10. 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?

First, we find the movies from each franchise in the dataset. Then for each movie we have:

H0: The median rating of each movie in the franchise is the same

Ha: The median rating is not consistent across movies of the franchise

Since this is rating data and for each franchise we have 3 or more movies, we compare the means using Kruskal-Wallis test by dropping nan entries for each movie during the comparison.

We get the following results:

Franchise	Number of movies	H	pval	Consistent (H0) / Inconsistent (Ha)
Star Wars	6	230.5842	8.0e-48	Inconsistent
Harry Potter	4	3.3312	0.3433	Consistent
The Matrix	3	48.3789	3.12e-11	Inconsistent
Indiana Jones	4	45.7942	6.27e-10	Inconsistent
Jurassic Park	3	46.5909	7.64e-11	Inconsistent
Pirates of the Caribbean	3	20.644	3.29e-05	Inconsistent
Toy Story	3	24.386	5.07e-06	Inconsistent
Batman	3	190.535	4.23e-42	Inconsistent