INTRO TO DATA SCIENCE

Project 1

Hypothesis Testing of movie ratings

11/07/2021

Name: Kewal Jani
Net-id: Kj2062
Email: kj2062@nyu.edu
NID: N13161679

1. Purpose and dataset discription

Purpose: In this project, we have to demonstrate the essential skills involved in hypothesis testing. To do so, we will use a real dataset that stems from a replication attempt of published research (Wallisch & Whritner, 2017).

Dataset description: This dataset features ratings data of 400 movies from 1097 research participants.
1st row: Headers (Movie titles/questions) – note that the indexing in this list is from 1 Row 2-1098: Responses from individual participants
Columns 1-400: These columns contain the ratings for the 400 movies (0 to 4, and missing)
Columns 401-421: These columns contain self-assessments on sensation seeking behaviors (1-5)
Columns 422-464: These columns contain responses to personality questions (1-5)
Columns 465-474: These columns contain self-reported movie experience ratings (1-5)
Column 475: Gender identity (1 = female, 2 = male, 3 = self-described)
Column 476: Only child (1 = yes, 0 = no, -1 = no response)
Column 477: Movies are best enjoyed alone (1 = yes, 0 = no, -1 = no response)
Note that we did most of the data munging for you already (e.g. Python interprets commas in a csv file as separators, so we removed all commas from movie titles), but you still need to handle missing data

Questions

1) ***Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular? [Hint: You can do a median-split of popularity to determine high vs. low popularity movies]***

Null Hypothesis : All Movies popular or less popular are rated Equally

To answer this question First I have to remove all the Nan values which I did using the element wise removal. After removing the Nan element-wise. I fetched all the data that has more reviews. (i.e., the movies with more reviews are considered popular as more people have reviewed it) and did the median split on the dataset.

After that I assembled all median of column of ratings of movies which are more popular in One column and the nonpopular one on another column and did the **u_test** As I split the data by median. so I got the value of **p = 1.9858517703414465e-34** which shows that it is **highly significant**. Hence, we can prove **our null hypothesis as false**. Which resembles that the mo vies that are more popular are rated higher than the movies that are less popular.

2) ***Are movies that are newer rated differently than movies that are older? [Hint: Do a median split of year of release to contrast movies in terms of whether they are old or new]***

Null hypothesis: old and new movies are rated the same

I split the data by years and found the median year as 1999. After that I assembled all the col umns of ratings of movies which are new in One column and the old one on another column and did **the u_test** As I split the data by median. so I got the value of **p = 0.40128144401686 45**. which shows that it is **not significant**. Hence, we can prove **our null hypothesis as true**. Which resembles that **old and new movie are rated the same**.

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

Null hypothesis: Shrek movie is rated equally by both genders.

Ans For this example I took the data set and divided the data set according to male and fema le columns and then removed the empty data which are of no use. After selecting the row ind ex I fetched data of ratings for shrek (2001) movie and did the **ks-test** as I have to deal with t he distribution. The result I got was 0**.056082040722863824~>0.005** which is not significan

t so I accepted that **the null hypothesis is correct**. That is the male and **female viewers do n ot rate the movie differently**.

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

Ans
Null hypothesis: Both Male and female rate movie equally.
For this example, I did the similar test as above. First Selected different row index which has male and female and divided rating data for each movie in two columns as I did then compare those two columns for each data set and whenever I found the **kstest** result less than 0,005 I considered that there is partiality in the movie rating. And then I took the ratio of the partiality result with total. I found out that **6.25% of movie were rated differently by male and female**.

5) ***Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?***
Ans :
Null Hypothesis: Both people who have only children and people with siblings enjoy movie equally
In this dataset I selected the column of ratings of the lion king movie and created a column o f people who watched the movie then created the column of the data for people with sibling and column for the people without sibling and did **kstest** on the two columns. I found the res ult as p= **0.15449987478607996** which is **not significant** hence, I approve the null hypothe sis that the **movies are enjoyed equally.**

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

Null Hypothesis: Both people who have only children and people with siblings enjoy movie equally

Ans for this example I did the similar test as above. First Selected different row index which has only child and not and then divided rating data for each movie in two columns. I did then compare those two columns for each data set and whenever I found the **ks-test** result less than 0,005 I considered that there is partiality in the movie rating. And then I took the ratio of the partiality result with total. I found out that only **0.75% of movie were rated differently**.

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?***

Null Hypothesis: Both people who like to watch movie socially and people who enjoy movie alone enjoy movie equally

Ans In this data I formed an index of rows who enjoys the movie alone and alternate column for the people who do not enjoy movie alone and then for the wolf of the wall street I

selected the rating column and distributed in the above 2 columns. Then I did ks-test among the movies and found out the value of **p = 0.49** which is not significant hence I accepted the null hypothesis. That is the wolf of the wall street is **enjoyed equally by the person who prefer to watch movies alone and the people who enjoys movie socially.**

8) ***What proportion of movies exhibit such a "social watching" effect?***

Null Hypothesis: Both people who like to watch movie socially and people who enjoy movie alone enjoy movie equally

Ans for this example I did the similar test as above. First Selected different row index which likes to watch alone with the people who watches movie socially and then divided rating data for each movie in two columns. I did then compare those two columns for each data set and whenever I found the **kstest** result less than 0,005 I considered that there is partiality in the movie rating. And then I took the ratio of the partiality result with total. I found out that 1.5% of movie were rated differently that **is 0.5% exhibits social watching effect**.

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

Null hypothesis: Distribution of ratings for Home alone and Finding Nemo movies are same.

Ans In this example I removed the NAN values or null values row wise and then formed 2 c olumns of rating each for Home alone and finding Nemo and then compare both the movies column. I got the value of **p= 1.3*e^-10** which **is significant** also I did a element wise remo val test for both the movies and got **the p value as 6.379381467525036e^-10**. Hence as both the **values are significant**, I can say that the **rating distribution for both the movies are di fferent.**

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? [Hint: You can use the keywords in quotation marks featured in this question to identify the movies that are part of each franchise]x***

Null hypothesis: All the movies franchise are consistent.

In this Example I selected all the movies given and for each movie I mined all the related movies or the sequels of the movie. After that I fixed the first part and did Kstest for the remaining movies. If I obtained the value of P<0.005 then it states that movies are not consistent. Which means that overall, the franchise is not consistent. I obtained that only **Harry potter and Pirates of the Carrabin are consistent movies franchise**.

11) **Extra** ***Weather horror movie like the ring(2002), The Exorcist (1973) is enjoyed when alone or with group ?***

Null hypothesis: movie is enjoyed equally in group or individually

Ans In this data I formed an index of rows who enjoys the movie alone and alternate column for the people who do not enjoy movie alone and then for the exorcist I selected the rating col umn and distributed in the above 2 columns. Then I did ks-test among the movies and found out the value of **p = 0.988391992186079, 0.8121477522294573.** which is not significant hen ce I accepted the null hypothesis. That is the ring and exorcist are **enjoyed equally by the p erson who prefer to watch movies alone and the people who enjoys movie socially**