Data Analysis Project 1
Haohai Pang, Jiasheng Ni
Oct 29th, 2023

Note: We learned in lecture 5 that assumptions of variances and normal distributions are often violated, but **the t-test is often considered as “robust” to violations of these assumptions.** So we chose t-tests to be our major test type in this analysis report. We assume that reducing movie ratings to the sample mean is reasonable because most movie sites use decimals to represent a movie's overall rating, which is the same as calculating sample means.

Question 1: Are movies that are more popular rated higher than movies that are less popular?

D: We made the null hypothesis “**H0: Popular movies didn’t rate higher than unpopular movies**” The alternative hypothesis would be “**H1: Popular movies rated higher than unpopular movies**”.
Then, we split 400 movies into popular and unpopular groups based on a median-split of popularity. We computed the average rating for each movie. Then, we calculated the inter-individual variance of the two groups to decide which test to be used. Finally, we performed the independent samples t-test.

Y: There are two sample means that are compared. The **inter-variances** are both <0.01, which are **small**. We consider the **variances to be similar**: both are nearly 0, so we use the **independent samples t-test**. This question is asking whether one group is rated higher than the other, so we follow the **one-tail** criteria to calculate the p value.

F: The median popularity is 197.5 among all 400 movies. We create two groups:

	Popular movies	Unpopular movies
Intra-Variability	0.0850024321309087	0.05357798836736424
Test-Statistics	-17.7560492698737	
P-Value	1.1348265138282423e-52	

A: Since p-value < alpha level = 0.005, we reject the null hypothesis, indicating that popular movies are rated significantly higher than unpopular ones.

Question 2: Are movies that are newer rated differently than movies that are older?

D: We made the null hypothesis “**H0: Newer movies didn’t rate differently than older movies**” and the alternative “**H1: Newer movies rated differently than older movies**”. We computed the mean ratings for each movie to represent the overall rating results. Next, we calculated the intra variance of two groups to decide which test to be used. Finally, we performed the independent t-test.

Y: We assume the ratings are continuous data, and thus means can be computed. We don't know the population parameters, and the **inter-individual variability is small and inter-group variance is similar, so we choose the independent t-test.**

F:

	Older Group(1000~1999)	Newer Group(1999~2016)
Intra-Variablity	0.118416102361980	0.1272191674221497
Test-Statistics	-1.7041093176937416	
P-Value	0.08914116542796356	

A: Since p-value > alpha level = 0.005, we fail to reject the null hypothesis, indicating there is no significant difference of the ratings between newer movies and older ones.

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

D: We made the null hypothesis "**H0: The movie 'Shrek (2001)' didn't rate differently by male and female viewers**" and the alternative hypothesis "**H1: The movie 'Shrek (2001)' rated differently by male and female viewers**". Then, we split the ratings of "Shrek" by female and male reviewers and calculate the variances. We use "nan policy = 'omit"' to **ignore missing values**. Finally, we performed the independent samples t-test.

Y: We assume that the ratings can be reduced to means. The **inter-variances** are both <1, which are **small**. Also the **variances are close** to each other, so we decided to use **the independent samples t-test.** However, we believe it is safe to verify our conclusion by assuming the variances (0.68 and 0.82) are different and perform **the Welch t-test** by specifying 'equal_var=False'. The p-value=0.248>alpha, our conclusion stays **the same**.

F:

	Male Viewers	Female viewers
Intra-Variability	0.6805843706777317	0.8218267169230881
Test-Statistics	-1.1016699726285888	
P-Value	0.27087511813734183	

A: Since p-value > alpha level = 0.005, we fail to reject the null hypothesis, indicating that the movie 'Shrek (2001)' didn't rate differently by male and female viewers

Question 4: What proportion of movies are rated differently by male and female viewers?

D: For each movie, we compared the mean ratings between male and female viewers (within each group, we drop N/A row-wise) and conducted the independent t-test to see if the differences between the average ratings were significantly different. Then accumulate the number of movies that have such significance.

Y: We assume that the ratings are continuous data and can be reduced to means. The inter-variability is small and similar (around 0~3 for female and male groups).

F: We found that about 11.5% of the movies are rated differently by male and female viewers.

A: Out of 400 movies, about 46 show the significance of the rating difference. So the proportion is about 11.5%.

Question 5: Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?

This question is like question 3. Similarly, we do the following steps:

D: We made the null hypothesis "**H0: people who are only children did not enjoy the movie The Lion King (1994)' more than people with siblings**" and the alternative "**H1: people who are only children enjoyed the movie 'The Lion King (1994)' more than people with siblings**". We created two groups of reviewers for this movie: people with siblings and people without siblings. Then, we calculated the inter variances of two groups to decide which test to use. We omitted missing values. Finally, we performed the **independent samples t-test.**

Y: We assume that the extent of enjoyment is directly reflected in the rating. The question is asking whether one group enjoys the movie more than the other, so we follow the **one-tail** criteria to calculate the p value by specifying the "alternatives" in the stats.ttest_ind function.

F:

	With siblings	Without siblings
Intra-Variability	0.5158031260392417	0.6666445916114792
Test-Statistics	2.053888996058986	
P-Value	0.9798664723686588	

A: Since p-value > alpha level = 0.005, we fail to reject the null hypothesis, indicating that people who are only children did not enjoy the movie The Lion King (1994)' more than people with siblings.

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

D: For each movie, we compared the mean ratings between viewers with siblins and those without (within each group, we drop N/A row-wise). and conducted the independent-t test to see if the differences between the average ratings were significantly different. Then accumulate the number of movies that have such significance.

Y: We assume that the ratings are continuous data and can be reduced to means. The inter-variability is small and similar (around 0~3 for viewers with child and without).

F: We found that about 2.5% of the movies are rated differently, where the p-value for the t-test is smaller than 0.005.

A: Out of 400 movies, about 10 show the significance of the rating difference. So the proportion is about 2.5%.

Question 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?

D: We first made the null hypothesis "H0: **people who like to watch movies socially did not enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone**" and the alternative hypothesis "H1: **people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone**." We split the rating of the movie into two groups: people who like to watch movies socially and people who don't. Then, we calculate the intra-variance of two groups to decide which test to use. We omitted missing values. Finally, we performed the **independent samples t-test.**

Y: We believe that the extent of enjoyment is directly reflected in the rating. Note that this question is asking whether one group enjoys the movie more than the other, so we should follow the one-tail criteria to calculate the p value by specifying the "alternatives='less'" in the stats.ttest_ind function.

F:

	Prefer socially	Prefer alone
Intra-Variability	0.8483271375464686	0.7567021083242457

Test-Statistics	-1.567873874504994
P-Value	0.9413054316716771

A: Since p-value > alpha level = 0.005, we fail to reject the null hypothesis, indicating that **people who like to watch movies socially enjoy ‘The Wolf of Wall Street (2013)’ more than those who prefer to watch them alone**.

Question 8: What proportion of movies exhibit such a “social watching” effect?

D: For each movie, we compared the mean ratings between viewers with “social watching effect” and those without (within each group, we drop N/A row-wise.) and conducted the independent-t test to see if the differences were significantly different. Then accumulate the number of movies that have such significance.

Y: We assume that the ratings are continuous data and can be reduced to means. The inter-variability is small and similar (still around 0~3 for viewers that exhibit “social watching” effect and those without).

F: We found that about 1.5% of the movies are rated differently where p-value for t-test are smaller than 0.005.

A: Out of 400 movies, about 6 movies show significance of rating difference. So the proportion is about 1.5%.

Question 9: Is the ratings distribution of ‘Home Alone (1990)’ different than that of ‘Finding Nemo (2003)’?

D: We conducted the KS-test to compare the rating distribution of the two movies.

Y: Since we are required to compare the rating distribution (assumed to be continuous data) of the two datasets, we can safely use the KS-test.

F: We found that the test-statistics is 0.15269080020897632, which is defined to be pointwise sup|f_1(x) - f_2(x)| (here f_1,f_2 are ecdf for two movie rating samples) and the p-value is 6.379397182836346e-10, which is less than 0.005.

A: From the p-value we get, we can safely reject the null hypothesis that the rating distribution of the two movies is the same and embrace the alternative hypothesis that the rating distributions are different.

Question 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]

D: For each franchise, we compared the mean ratings between different movies (within each group, we drop N/A row-wise.) and conducted the one-way ANOVA test to see if the means are different (showing inconsistency). The null hypothesis should be: All means ratings of the movies from the same franchise are the same and the alternative one should be: at least one of these average ratings differ. Finally, accumulate the number of franchises that have significance.

Y: We assume that the ratings are continuous data and can be reduced to means. We want to compare between multiple groups (multiple movies > 3 by our data) within the same franchise, so the one-way ANOVA test suits best.

F: We found that about 75% of the franchises(marked in yellow) are rated inconsistently where p-value for ANOVA test are smaller than 0.005.

	F-score	P-value
Star Wars	39.029939613074056	2.399595163532992e-38
Harry Potter	1.4456904473285563	0.2275340290918136
The Matrix	18.59281511303809	1.2957225925356723e-08
Indiana Jones	19.050958699528884	5.20425425762115e-12
Jurassic Park	22.163615231952207	3.542127514286409e-10
Pirates of the Carribean	3.4465950041304323	0.032079328032699014
Toy Story	7.5881445425788305	0.0005193828629536133
Batman	43.62587891516757	1.6410731510652517e-18

A: Out of 8 franchises, 6 show the significance of rating inconsistency. To be specific, except for the Harry Potter and Pirates of the Caribbean franchise, all the other movies show rating inconsistency.

Extra Credit: Tell us something interesting and true (supported by a significance test of some kind) about the movies in this dataset that is not already covered by the questions above [for 5% of the grade score].

Fact: In general, people of different genders (male, female, others) did not rate movies differently.

D: We first made the null hypothesis to be “**H0: In general, people of different genders did not rate movies differently”** and the alternative hypothesis “**H1: In general, people of different genders rated movies differently”.** Next, we calculated the average rating a person gives to all movies this person has rated. Then, we **created three groups** of reviewers: [male, female, and others], consisting of the average rating of each user. We computed the intra variances just for fun. There are three groups that are compared. So we performed the **ANOVA Welch t-test.** Results are shown as follows:

	female	**male**	**other**
Intra-variance	0.2204158439117208	0.2666415455767475	0.16607817233199895
f_statistic	1.677625761642752		
p_value	0.18730800328635697		

Since p-value > alpha level = 0.005, we fail to reject the null hypothesis, indicating that n general, people of different genders (male, female, others) did not rate movies differently.

Attachments:

Code Link: https://github.com/AlexMan2000/NYU-Master-Program/blob/master/DS-GA-1001/Data_Analysis_Project/Data_Analysis_Project_1.ipynb